AD_____

Award Number: DAMD17-02-1-0214

TITLE: Digital Mammography: Development of an Advanced Computer-Aided System for Breast Cancer Detection

PRINCIPAL INVESTIGATOR: Heang Ping Chan, Ph.D.

CONTRACTING ORGANIZATION: University of Michigan
Ann Arbor, Michigan 48109-1274

REPORT DATE: May 2004

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 2004	Annual (1 May 03-30 Apr 04)

4. TITLE AND SUBTITLE
Digital Mammography: Development of an Advanced Computer-Aided System for Breast Cancer Detection

5. FUNDING NUMBERS
DAMD17-02-1-0214

6. AUTHOR(S)
Heang Ping Chan, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Michigan
Ann Arbor, Michigan 48109-1274

E-Mail: chanhp@umich.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

The goal of the project is to develop computer-aided diagnosis (CAD) methods and systems for mammography using advanced computer vision techniques and image information fusion from multiple mammograms to improve lesion detection and characterization. When fully developed, the CAD system can assist radiologists in mammographic interpretation.

During this project year, we have performed the following tasks: (1) collected databases of digital mammograms (DMs) and digitized film mammograms (DFMs) for development of the CAD systems, (2) conducted a study to compare the percent dense area manually segmented by experienced radiologists on DMs and DFMs, (3) developed new image enhancement techniques and new false-positive reduction methods for mass detection, and conducted studies to compare the accuracy of mass detection by the CAD systems for DMs and DFMs using FROC analysis, (4) developed automated method for nipple detection on mammograms as a basis of multiple image fusion analysis for CAD systems, and (5) compared the accuracy for classification of malignant and benign breast masses using single-view and fused two-view information on mammograms by computer, and evaluated the effects of CAD on experienced radiologists' characterization of malignant and benign breast masses in two-view temporal pairs of mammograms.

In summary, we have investigated a number of areas in CAD of mammographic lesions and evaluated the new techniques for both DMs and DFMs. We have made progress in the six tasks proposed in the project. We have found that our new computer-vision techniques and two-view information fusion approach can improve the performance of the CAD systems. We will continue the development of the CAD systems for DMs and DFMs in the coming years.

14. SUBJECT TERMS
Breast Cancer, Digital Mammography, Computer-Aided Diagnosis, Breast Cancer Diagnosis

15. NUMBER OF PAGES
52

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

(3) Table of Contents

(4) Introduction

Computer-aided diagnosis (CAD) has been shown to be useful as a second opinion to radiologists for breast cancer detection on mammograms. All current CAD systems have been developed for digitized screen-film mammograms (DFM). With the recent advent of full field digital mammography (FFDM) systems, it is important to develop CAD systems specifically designed for direct digital mammograms (DMs) in order to fully exploit the advantages of FFDM. Although many computer vision techniques developed for digitized films may be used for DMs, proper adaptation and extensive training of the current algorithms for the new type of images will be required. More importantly, new techniques still need to be developed to further improve the current algorithms for DFMs as well as for adapting to FFDM.

The goal of the proposed research is to develop a CAD system for breast cancer diagnosis using advanced computer vision techniques. The proposed CAD system will assist radiologists with detection and classification of breast lesions. Previous CAD methods for lesion detection and characterization are generally based on image features extracted from a single view. Our proposed approach is based on two steps: the first step uses single view detection to identify lesion candidates on individual mammograms, the second step is to fuse image information from multiple views to reduce false positives and thus to improve the overall accuracy. Although the main goal of this project is to develop a CAD system for DMs, we plan to extend the CAD development to DFMs for the following reasons: (1) digital mammography only became available in the last few years, multiple-view film mammograms with breast lesions are more commonly available in existing patient files, and (2) screen-film mammography will still be the main modality for breast cancer screening in the near future. Therefore, we will first develop the multiple-view correlation techniques for the CAD system of the DFMs. These new techniques will then be adapted to the CAD system for DMs. We believe that this approach is more efficient and we will obtain a CAD system for DMs as well as improve the CAD system for DFMs.

The following specific aims will be addressed: (1) Collection of databases of both DMs and DFMs and design of a database management system. (2) Improvement of single-view computer vision techniques for mass detection and classification in DFMs. (3) Improvement of single-view computer vision techniques for microcalcification detection and classification in DFMs. (4) Development of methods for correlation of image information from two-view DFMs. (5) Comparison of the detection and classification accuracy of the multiple-view fusion CAD system with the performance of the single-view CAD system by receiver operating characteristic (ROC) and FROC analyses. (6) Adaptation of the computer vision techniques to the CAD system for DMs. (7) Adaptation of the multiple-view fusion methods to the CAD system for DMs.

We will develop novel regional registration methods for identifying corresponding lesions on craniocaudal (CC) and mediolateral oblique (MLO) views. The multiple image information will be fused with specially designed correspondence classifiers or fuzzy classification to reduce false positives and to improve lesion detection sensitivity. Multiple-view features of a lesion will be merged using neural networks or other classifiers for classification of malignant and benign lesions. In addition, new computer vision techniques will be developed in each of the four areas to improve the current methods. The techniques will be first developed for DFMs. The algorithms for DFMs will then be adapted to DMs, taking into account the differences in the imaging characteristics between DMs and DFMs. Databases of DFMs and DMs will be collected from our patient population with IRB approved protocol and extensive training and independent testing of the new CAD system will be

performed. The test performance of the multiple-image correlation CAD algorithms for detection and characterization of lesions on DFMs will be compared with the one-view approach on DFMs as well as the performances of CAD systems for DMs using ROC methodology.

DM or DFM not only has the potential to detect breast cancer in an early stage, it will also facilitate consultation via teleradiology in remote or rural regions where expert mammographers may not be readily available. An effective CAD system will be particularly useful for providing an additional on-site or remote second opinion. This will be highly relevant to women in the military, especially when they are stationed in remote areas. DM in combination with CAD will fully utilize the potential of mammography to improve the health care of women both in the military and in the general population.

(5) Body

This is the second year annual report of our project. In the project period (5/1/03-4/30/04), we have extended our investigations to both the CAD systems for DMs and DFMs, and performed a number of studies to develop the CAD system for breast cancer diagnosis. A summary of some of the important accomplishments follows.

(A) Collection of databases of digital mammograms and digitized film mammograms

We continue to collect the database of digital mammograms (DMs) with mammographic masses or clustered microcalcifications for the development of our computer-aided diagnosis (CAD) algorithms. We have collected about 140 cases. The patients were diagnosed with in their mammograms during their normal clinical care, either by routine screening or by referral to our breast imaging clinic for evaluation. Most of the cases contained both DMs and screen-film mammograms. The digital mammograms were acquired with a GE Senographe 2000D full field digital mammography (FFDM) system. The pixel size of the system is 100 μm X 100 μm. The gray level resolution of the system is 14 bits for the raw images and 12 bits for the processed images. After acquisition, the digital image files are transmitted to the Siemens Archive which is the PACS system used in our department for storage of all clinical digital images.

With Institutional Review Board (IRB) approval, we retrieved the DMs from the Siemens Archive to our laboratory and digitized the film mammograms from the same patient. The film mammograms were digitized with a Lumiscan 85 laser scanner at a pixel size of 50 μm X 50 μm and a 12 bit gray level. We have developed a database management program based on Microsoft Access to process the images downloaded to our system. For each mammogram file, all patient identifiers are first removed from the image header. The patient name is replaced with a code number. The image is then named by the code number, the view (craniocaudal, mediolateral oblique, or mediolateral), and the exam year. A record is generated in the database file for each image. The record keeps the code number, the lesion type, the view, and the exam date information for each case. When the pathology of the case is available, the malignant or benign information of the lesion is also entered. Each case in the database will be read by an experienced MQSA radiologist to mark the lesion location. For microcalcification cases, the radiologist measures the diameter of the cluster, and provides description of its distribution,

morphology, and visibility of the microcalcifications. For mass cases, the radiologist measures the diameter of the mass, and provides description of its margin, shape, spiculated or non-spiculated, the visibility, and the density of the mass relative to that of the parenchyma. For all cases, the radiologist also provides BI-RADS description of the breast density and estimates the likelihood of malignancy of the lesion. These descriptions are entered into the database for each case as a reference for future analysis.

(B) Comparison of density segmentation on digitized film mammograms and digital mammograms

Mammographic sensitivity is limited by the breast density. Dense fibroglandular tissue appears as low optical density regions on mammograms. If masses or clustered microcalcifications overlap with the dense tissue, the radiographic contrast between the lesion and the normal tissue will be low and the detectability of the lesion will be reduced. One of the expected advantages of DM over film mammograms is that the higher contrast sensitivity of digital detector and the image processing applied to the DMs may improve the mammographic sensitivity and thus reduce the false negative detection rate of mammography. We have performed a study to compare the breast density estimated on pairs of digital mammogram (DM) and (DFM) obtained from the same patients. This study may provide useful insight on the relative performance of DMs and DFMs for radiologist's interpretation and computerized image analysis. We reported the preliminary results of this study in last year's annual report. The study was completed during the current year and was presented at the Radiological Society of North America (RSNA) Annual Meeting in the November of 2003. The study is summarized below.

Methods:

One hundred ninety-eight pairs of DM and DFM (99 CC views and 99 MLO views) were collected with IRB approval from 99 patients. The time interval between the DM and DFM ranged from 0 to 118 days (median=26.3 days). The DFMs were acquired with GE DMR systems and the DMs were acquired with the GE Senographe 2000D system. Both the DMs and the DFMs were acquired with automated exposure techniques that selected the appropriate target, filter, and kVp. The breast boundaries on the DMs and DFMs were detected automatically by the computer. The mammograms were displayed on a workstation with a graphical user interface (Fig. 1) that allowed the radiologist to perform interactive thresholding of the gray level histograms to segment the dense region from the fatty region. The DMs and DFMs were segmented independently in separate sessions so that the reader could not compare the density of the corresponding DM and DFM. The mammographic density was estimated as the percent dense area relative to the breast area. For the MLO views, the pectoral muscle was excluded from the breast area calculation. Five MQSA radiologists participated as readers and segmented the breast density independently with interactive thresholding.

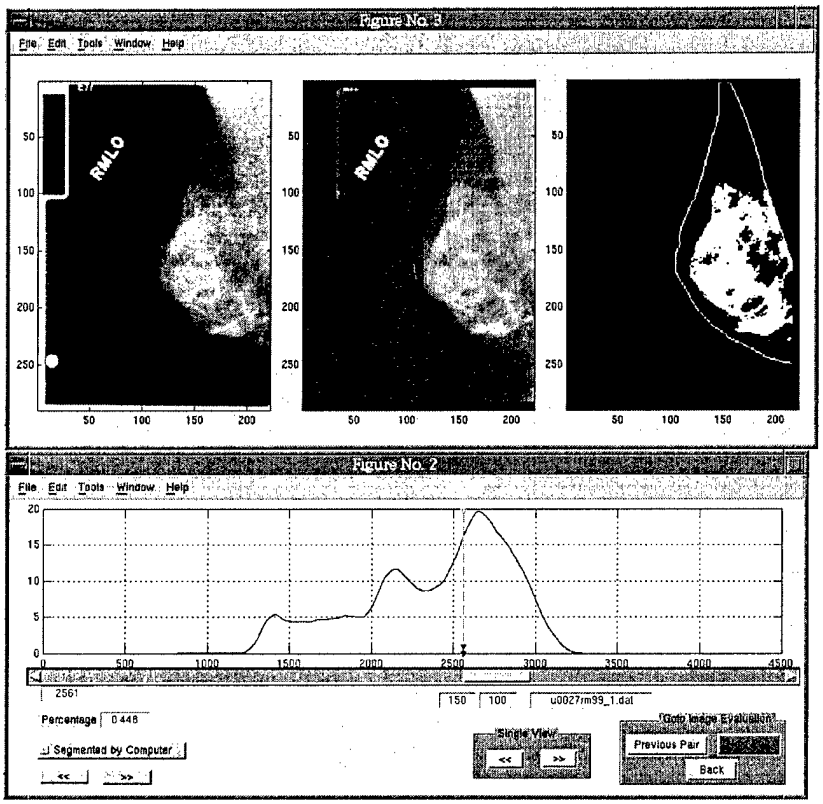


Fig. 1. An example of density segmentation on digital mammograms. Left: original mammogram. Middle: processed image. Right: segmentation result of breast density. Lower: Gray level histogram of processed image with a manually selected threshold.

Results:

We analyzed the x-ray imaging techniques selected automatically by the screen-film mammography systems and the digital mammography system. The distribution of the target/filter combination for the CC and the MLO views are compared separately in Fig. 2. For the DM system, the majority of the mammograms were acquired with the Rh/Rh combination. For the screen-film system, the majority of the mammograms were acquired with the Mo/Mo combination. The difference in the kV settings between the DM and the screen-film mammograms are shown in Fig. 3. Most of the DMs were acquired with a higher kV than that for the screen-film mammograms. The average difference in kV was 2.4 kV and 2.5 kV, respectively, for the CC view and the MLO view.

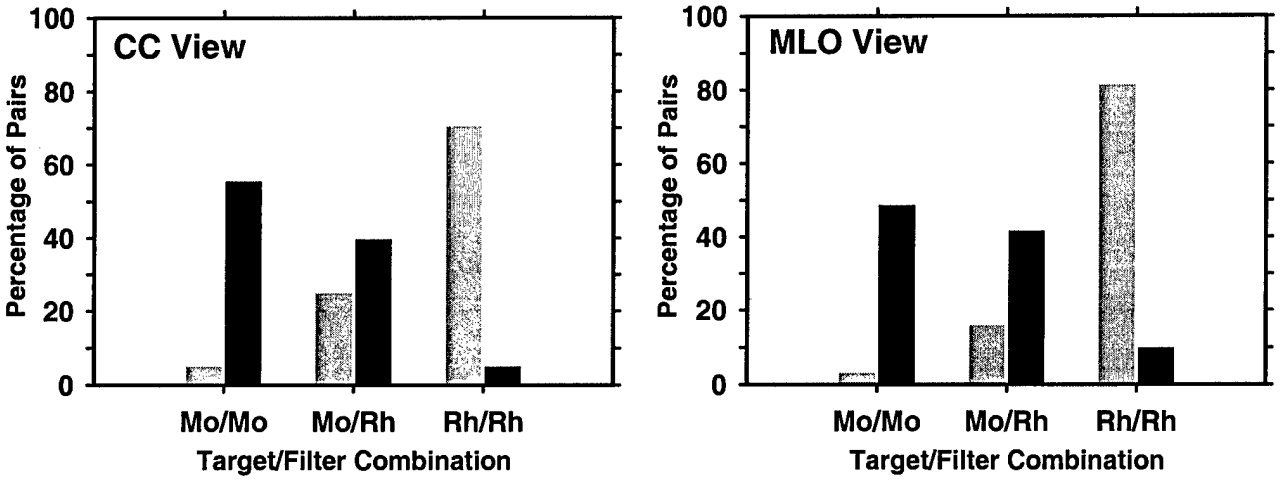


Fig. 2. Comparison of the target/filter combination selected by the x-ray system for acquisition of the digital and digitized mammograms. Left: CC view; Right: MLO view.

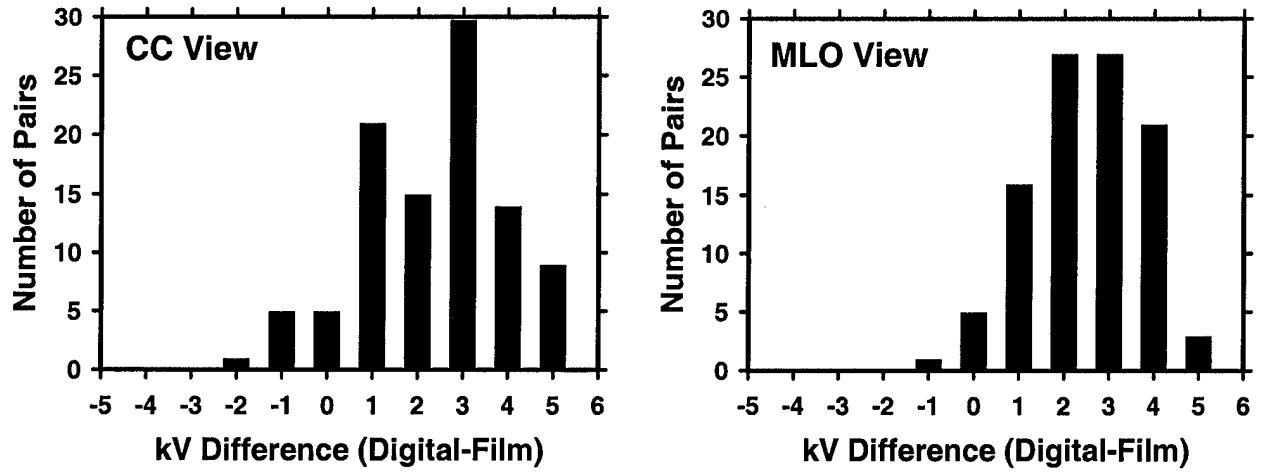


Fig. 3. The difference in kV selected by the x-ray system for acquisition of the digital and digitized mammograms. Left: CC view; Right: MLO view.

An example of density segmentation with interactive thresholding using the graphical user interface is shown in Fig. 1. Fig. 4 shows the distribution of the difference in the % breast density segmented from the DFM compared to that of the DM for the same breast and the same view. The majority of the differences are positive. Fig. 5 shows the distribution of the ratio of the % breast density of the DFM to that of the corresponding DM for the same breast and the same view. The majority of the ratios are greater than 1. Table 1 summarizes the mean difference in the % breast density and the mean ratio of the % breast density for the corresponding DM and DFM for the same breast and the same view for the 5 radiologists. The mean differences for four of the five radiologists were greater than zero and the mean ratios were greater than 1, indicating that the DFM was perceived by the radiologists as more dense than the DM. One of the five radiologists had a mean difference smaller than zero and a mean ratio of smaller than, indicating that this radiologist perceived the breast density as more dense in the DMs. The mean difference and the mean ratio over all 5 radiologists were positive and greater than 1, respectively.

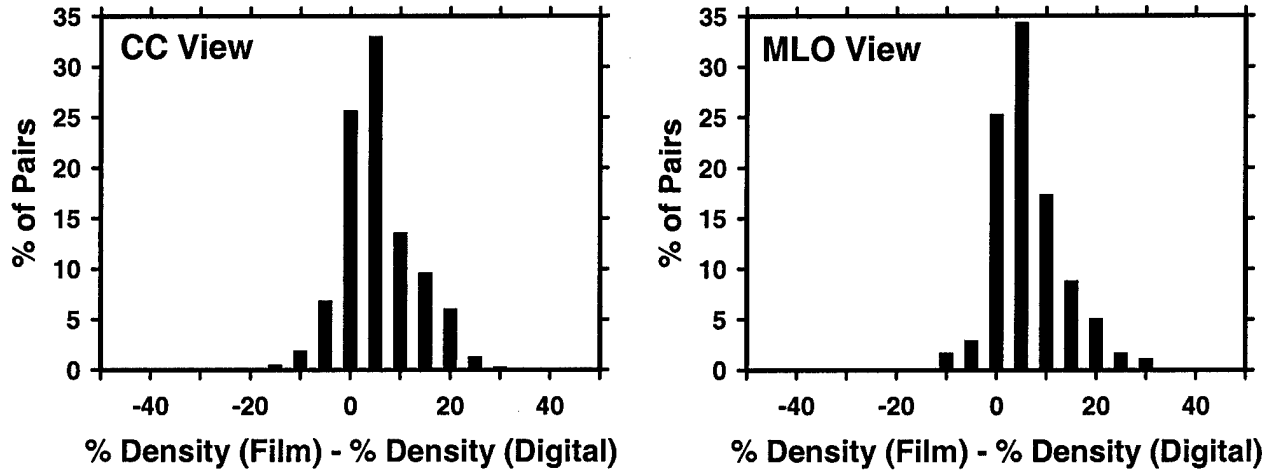


Fig. 4. The difference in % breast density between the corresponding DFM and DM of the same breast and the same view. Left: CC view; Right: MLO view.

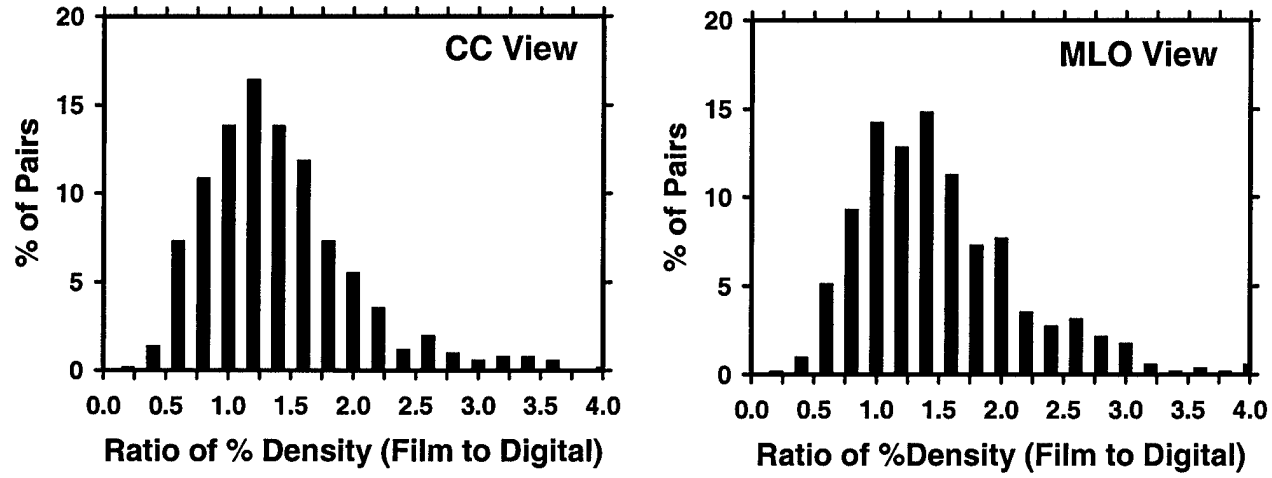


Fig. 5. The ratio of % breast density between the corresponding DFM and DM of the same breast and the same view. Left: CC view; Right: MLO view.

Conclusion:

The mammographic density was significantly higher on film mammograms than on digital mammograms, as segmented by 4 of the 5 radiologists. On average, the perceived breast density was about 30% higher on film mammograms than on digital mammograms. The difference in the perceived density may be caused by the harder beam quality used and the image processing applied to the DMs. The lower density on DMs may improve the mammographic sensitivity for lesion detection on dense breasts. However, radiologists often compared current and prior mammograms for detection of developing focal density that may be an early sign of breast cancer. If a patient has DFMs in prior exams and the new exam is performed with DM, the significant reduction in mammographic density due to the change in imaging techniques may reduce the sensitivity of detecting developing new

focal density. The radiologists will have to take into consideration the differences between the properties of DFMs and DMs when reading these cases.

Table 1. Comparison of the mean % breast density manually segmented by 5 radiologists from corresponding DFM and DM of the same breast and the same view.

Radiologist	Mean Difference in % Density *(Film – Digital)*	Mean Ratio of % Density *(Film / Digital)*
Rad 1	4.92	1.35
Rad 2	5.50	1.53
Rad 3	4.63	1.39
Rad 4	4.66	1.35
Rad 5	-2.01	0.88
Average (all)	3.54	1.30
Average (#1 - #4)	4.93	1.40

(C) Computer-aided diagnosis system for mass detection: comparison of performance on digital mammograms and digitized film mammograms

We have been investigating methods for improvement of the CAD system for detection of masses on DFMs as well as adapting the CAD system to mass detection on DMs. In the annual report of 2003, we have reported a preliminary study of comparing the performance of the two systems on pairs of DM and DFM images obtained from the same patients. In this project year, we have evaluated new image enhancement techniques and false positive (FP) reduction methods and compared the performance of the two improved systems with a larger data set. The results were reported in the RSNA meeting of 2003 and the PRMRP Investigators Meeting in April of this year. The study is summarized below.

Methods:

Our CAD system consisted of four steps: image enhancement, clustering-based region growing and local refinement, extraction of morphological and texture features, and rule-based and linear classification for FP reduction. Previously, image enhancement was achieved by a density-weighted contrast enhancement (DWCE) filter. In this study, the DWCE filter was replaced by a gradient field analysis method that located mass candidates based on the locations where strong gradient converged radially towards a point. New gradient field feature and morphological features were incorporated into the linear classifier in the FP reduction step. The simplex optimization procedure was used in the stepwise feature selection process to select the most effective features for classification of true masses and normal breast tissues. A schematic of the mass detection CAD system for DMs is shown in Fig. 6. The system for DFMs is similar except that it does not need preprocessing with the multiscale enhancement. The Laplacian Pyramid preprocessing for DMs was described in last year's annual report.

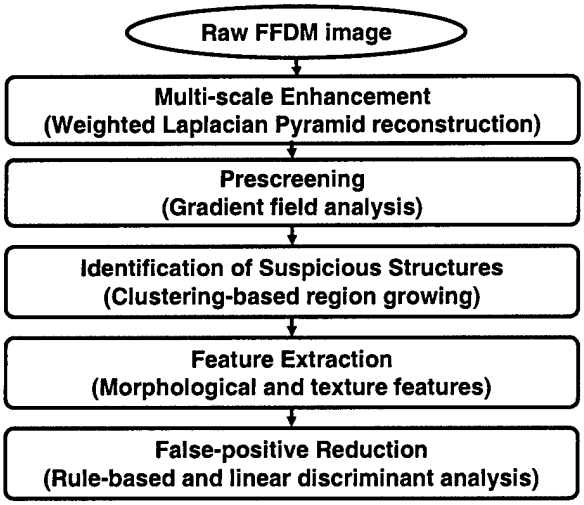


Fig. 6. Schematic for mass detection in full-field digital mammograms.

We used a data set of 100 cases containing two-view DMs acquired with a GE FFDM system and the corresponding DFMs of the same views for the same breast. The data set contained 102 masses. The true locations of the masses were identified by an experienced MQSA radiologist. The data set was split into a training set and a test set, with 50 cases in each set. The new mass detection CAD system was trained separately for DMs and DFMs. The CAD system trained with DFMs was applied to the test set of DFMs, and that trained with DMs was applied to the test set of DMs. The performances of the CAD systems for the DMs and the DFMs were compared by the free response receiver operating characteristic (FROC) analysis.

Results:

The FROC curves for the training and test sets are plotted in Fig. 7 and Fig. 8, respectively. The FROC curve shows the detection sensitivity (TPF) as a function of FPs per image and thus shows the trade-off between sensitivity and specificity of a detection algorithm. The sensitivity was estimated on single-view (image-based) mammogram as well as on two-view (case-based) mammograms. One-view scoring counts the mass on each mammogram independently. On the other hand, two-view scoring counts a TP if the mass is detected at least on one of the two views and the total number of masses is the total number of two-view pairs. Two-view scoring assumes that a radiologist will not overlook the mass as long as the CAD system marks the mass in at least one view. Two-view scoring generally shows a higher sensitivity because a mass is counted as TP even if it is missed in one of the two views. The FP rates for the one view and two-view detection are also tabulated in Table 2 and Table 3 at several sensitivity values for the training set and test set, respectively. It can be seen that the FROC curves for the DM and the DFM are very close for the training set. The FROC curves for the DMs are slightly higher than those for the DFMs in the test set. The FP rates are 2.0 and 2.1 per image for DMs and DFMs, respectively at 90% sensitivity and 1.7 and 1.9 per image, respectively, at 85% sensitivity.

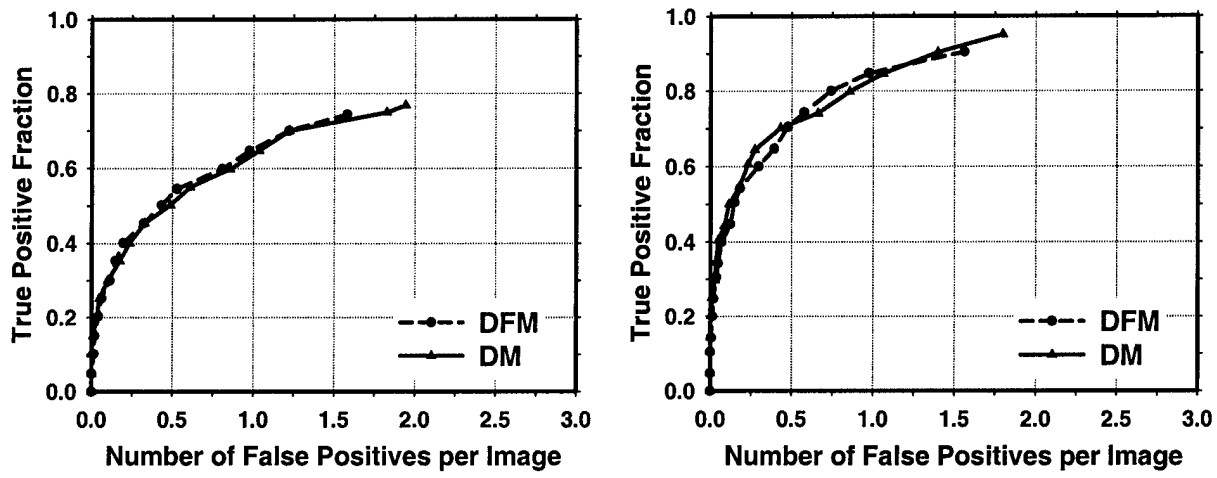


Fig. 7. FROC curves comparing the performance of computer detection on DMs and DFMs. The curves are obtained from the training set and the sensitivity is determined by detection in Left: single-view mammograms. Right: two-view mammograms.

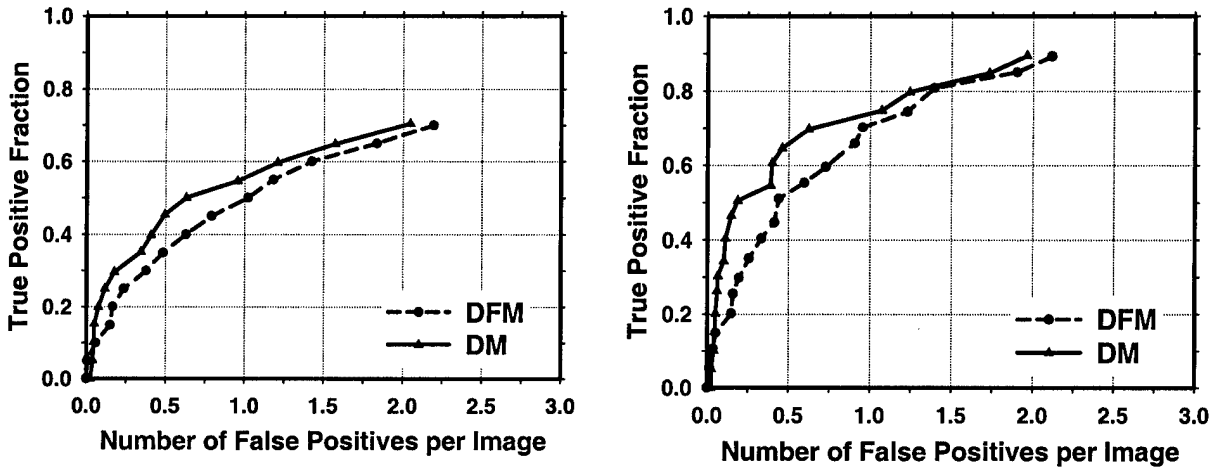


Fig. 8. FROC curves comparing the performance of computer detection on DMs and DFMs. The curves are obtained from the test set and the sensitivity is determined by detection in: Left: single-view mammograms. Right: two-view mammograms.

Table 2. Comparison of detection accuracy in full field digital mammogram (DM) and digitized film mammograms (DFM) for the training set. TPF=true positive fraction, FP=false positive

Image-based			Case-based		
TPF	FPs/image		TPF	FPs/image	
	DM	DFM		DM	DFM
75%	1.8	1.6	90%	1.4	1.6
70%	1.2	1.2	85%	1.0	1.0
65%	1.0	1.0	80%	0.8	0.7

Table 3. Comparison of detection accuracy in full field digital mammogram (DM) and digitized film mammograms (DFM) for the test set. TPF=true positive fraction, FP=false positive

Image-based			Case-based		
TPF	FPs/image		TPF	FPs/image	
	DM	DFM		DM	DFM
75%			90%	2.0	2.1
70%	2.0	2.2	85%	1.7	1.9
65%	1.6	1.8	80%	1.2	1.4

Conclusion:

After training with case samples from each modality, our mass detection CAD scheme can be useful for detecting masses on both DMs and DFMs. Further study is underway to improve the various stages of each of the mass detection systems based on the properties of the DM and DFM images, respectively.

(D) Computer-aided diagnosis on mammograms using multiple image analysis: computerized nipple identification

Correlation of information from multiple-view mammograms (e.g., MLO and CC views, bilateral views, or current and prior mammograms) can improve the performance of breast cancer diagnosis by radiologists or by computer. Nipple is a reliable and stable landmark on mammograms for the registration of multiple mammograms. However, accurate identification of nipple location on mammograms is challenging because of the variations in image quality and in the nipple projections, resulting in some nipples nearly invisible on the mammograms. In this study, a computerized method was developed to automatically identify nipple location on mammograms.

Methods:

We developed a two-stage method for nipple detection as shown in Fig. 9. The nipple location was identified using the gray level information around the nipple, geometric characteristics of nipple shapes, and the texture features of glandular tissue or ducts converging towards the nipple. The breast boundary was first obtained using a gradient-based boundary tracking algorithm, then the gray level profiles along the inside and outside of the boundary were extracted. A geometric convergence analysis was used to limit the nipple search to a region of the breast boundary. At the first stage, a rule-based method was designed to identify the nipple location by detecting significant changes of intensity along the gray level profiles inside and outside the breast boundary and the changes in the boundary direction. At the second stage, a texture orientation-field analysis was developed to estimate the nipple location based on the convergence of the texture pattern of glandular tissue or ducts towards the nipple. The nipple location was finally determined from the detected nipple candidates by a rule-based confidence analysis

We randomly selected 377 and 367 DFMs for training and testing the nipple detection algorithm. The nipple location identified by two experienced radiologists was used as the ground truth. In the training data set, 301 nipples could be identified positively and were referred to as visible nipples, 76 nipples could not be identified positively and were referred to as invisible nipples. In the test data set, 298 and 69 of the nipples were identified as visible and invisible, respectively. The nipple locations of the invisible group were estimated by the radiologists for comparison with the computer estimates.

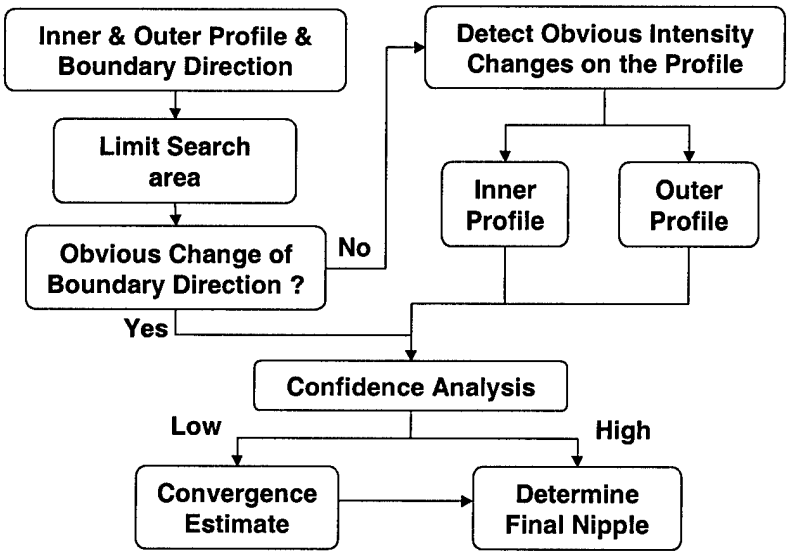


Fig. 9. Schematic of the automated nipple search method

Results:

The detection accuracy for training and testing of the algorithms is summarized in Table 4. In the training set, the computerized method could detect 89.37% (269/301) of the

visible nipples and 69.74% (53/76) of the invisible nipples within 1 cm of the truth. In the test data set, 92.28% (275/298) of the visible nipples and 53.62% (37/69) of the invisible nipples were identified within 1 cm of the truth.

Table 4. Performance of the automated nipple detection program. The nipple detection accuracy is quantified as the percentage of images in which the detected nipple location is within 1 cm to the ground truth.

		Number of Images	Rule-based method	Rule-based method with texture analysis
Training set	Visible	301	82.39% (248/301)	89.37% (269/301)
	Invisible	76	65.79% (50/76)	69.74% (53/76)
	all	377	79.05% (298/377)	85.41% (322/377)
Test set	Visible	298	89.93% (268/298)	92.28% (275/298)
	Invisible	69	47.83% (33/69)	53.62% (37/69)
	All	367	82.02% (301/367)	85.01% (312/367)

Conclusion:

Accurate identification of nipple location on mammograms is challenging because of the variations in image quality and in the nipple projections, especially for the nipples that are nearly invisible on the mammograms. In this work, we developed a two-stage computerized nipple identification method to detect or estimate the nipple location. The results demonstrate that the visible nipples can be accurately detected by our computerized image analysis method. The nipple location can be reasonably estimated even if it is invisible. Automatic nipple identification will provide the foundation for multiple image analysis in CAD.

(E) Evaluation of two-view fusion techniques and the effects of computer-aided diagnosis on radiologists' characterization of malignant and benign breast masses in two-view temporal pairs of mammograms

We are developing an automated CAD program to classify masses as malignant or benign based on interval change information on serial mammograms. The classifier was initially developed with single-view mammograms. In this project year, we compared fusion methods to combine the single-view classifier output scores to two-view (CC and MLO views) scores and studied if two-view classification could improve classification accuracy. We also conducted observer performance experiments with ROC methodology to evaluate the effects of CAD on radiologists' estimates of the likelihood of malignancy of masses.

Methods:

We have designed a computerized method to analyze the current and prior information for a temporal pair. In each region of interest (ROI) containing the current or prior mass, the segmentation of the mass from the surrounding background tissue was carried out in two stages. K-mean clustering was first used to extract an initial contour. The initial contour was then used as the starting point for the active contour model, which allowed more accurate and refined segmentation of the mass boundary. From each automatically segmented mass, a total of 35 features, including 20 run-length statistics (RLS) texture features, 12 morphological and 3 spiculation features, were extracted from each ROI. Additionally, difference features were derived by subtracting a feature of the prior mass from the corresponding feature of the current mass. Therefore, 35 difference features were derived from the 20 RLS texture, 12 morphological and 3 spiculation features.

A total of 300 mammograms, containing CC and MLO views and serial exams, was obtained from the files of 68 patients, from which 90 two-view temporal pairs of the masses (47 malignant and 43 benign) were formed. The classifier was designed based on the single-view CC and MLO temporal pairs. A "leave-one-case-out" resampling method was used for training and testing the classifier. A test classifier score was obtained for each single-view CC or MLO temporal pair. The score for a two-view temporal pair was then derived by merging the scores of the corresponding CC and MLO single-view temporal pairs of the same mass. We compared three fusion methods in this study: maximum, minimum, and average, and found that averaging provided the highest classification accuracy. The average score was therefore obtained from the single-view scores, and this merged score was used as an estimate of the relative malignancy rating from the two views of the mass by the classifier. The average score was linearly transformed to a scale of 1 to 10 and the transformed rating was presented to the observers in the observer study. Eight MQSA radiologists and 2 breast imaging fellows participated as observers in this study. The 90 two-view temporal pairs of masses were divided into two case groups. Each observer read the 180 cases (90 cases X two reading conditions: with and without CAD) in two reading sessions, separated by at least one month. In each session, one case group was read using the without-CAD mode and the other using the with-CAD mode. The order of the two reading conditions was switched between the two reading sessions and the order of the cases within each case group was randomized for each observer. The orders of the case groups and the reading conditions were arranged in a counter-balanced design such that no case group or reading condition would be read first more often than the other when averaged over all observers.

Results:

The results indicated that two-view fusion of the classifier scores improved the classification accuracy (A_z) from 0.86 for the 180 single-view temporal pairs to 0.90 for the corresponding 90 two-view temporal pairs. For the observer study, the observers' estimated likelihood of malignancy ratings were analyzed by the Dorfman-Berbaum-Metz (DBM) multi-reader multi-case (MRMC) methodology (1). The ROC curve was derived from a maximum likelihood estimation of the binormal distributions fitted to the

observers' rating data and the area under the ROC curve, A_z, was calculated. The statistical significance of the difference between the studied modalities was estimated by the DBM method and by the Student's two-tailed paired t-test for observer-specific paired data. The A_z values and the partial area index above a sensitivity of 0.9, $A_z^{(0.9)}$, for the characterization of the masses in the two reading modes: without-CAD and with-CAD by the 10 radiologists are tabulated in Table 5. The improvement in A_z without CAD to with CAD was statistically significant (p=0.031, Student's paired t-test; p=0.046, DBM method). The partial area index for the reading with CAD also improved from that without CAD, but the improvement did not achieve statistical significance.

Conclusion:

Two-view information fusion can improve the classification accuracy for breast masses. From the results of our observer study that evaluated the effects of CAD on radiologists' characterization of masses, we found that CAD using interval change analysis on two-view serial mammograms can significantly improve radiologists' classification accuracy of masses and thereby may improve the accuracy of biopsy recommendations. The reduction in unnecessary biopsy will reduce health care costs and patient anxiety.

Table 5. The area under ROC curve, A_z, and the partial area index above a sensitivity of 0.9, $A_z^{(0.9)}$, for the characterization of the masses in the two reading modes: without-CAD and with-CAD by the 10 radiologists.

Radiologist No.	A_z Without-CAD	A_z With CAD	$A_z^{(0.9)}$ Without-CAD	$A_z^{(0.9)}$ With CAD
1	0.75 ± 0.05	0.88 ± 0.04	0.13	0.63
2	0.86 ± 0.04	0.86 ± 0.04	0.41	0.52
3	0.80 ± 0.05	0.80 ± 0.05	0.17	0.30
4	0.88 ± 0.04	0.92 ± 0.03	0.46	0.73
5	0.74 ± 0.05	0.84 ± 0.04	0.20	0.48
6	0.86 ± 0.04	0.87 ± 0.04	0.48	0.31
7	0.85 ± 0.04	0.85 ± 0.04	0.52	0.56
8	0.88 ± 0.04	0.86 ± 0.04	0.54	0.46
9	0.78 ± 0.05	0.83 ± 0.04	0.15	0.18
10	0.77 ± 0.05	0.84 ± 0.04	0.22	0.21
Average	0.83	0.87	0.35	0.49

(6) Key Research Accomplishments

- Continue collection of a database of digital mammograms and digitized film mammograms for development of the CAD algorithms for both digital mammography and film mammography ------ (Task 1)

- Complete the observer study for comparison of density segmentation on digitized screen-film mammograms and digital mammograms. Understand the differences between the imaging techniques and the image properties of digital mammograms and digitized mammograms ------- (Task 1, Task 2, Task 3, Task 5)

- Continue the development of CAD mass detection systems for digital mammograms and digitized film mammograms. Improve the prescreening methods and feature classifiers in both systems. Perform a study to compare the performance of the two improved CAD systems for mass detection on digital mammograms and digitized film mammograms by FROC analysis ------- (Task 2)

- Develop automated nipple detection method that will serve as a basis for multiple image fusion analysis for an advanced CAD system ------ (Task 4, Task 5, Task 6)

- Develop a computerized method for classification of malignant and benign masses using interval change analysis of two-view mammograms. Investigate the effects of CAD on radiologists' performance on classification of masses ---- -- (Task 2, Task 4, Task 6)

(7) Reportable Outcomes

As a result of the support by the PRMRP grant, we have conducted studies in CAD for mammography and published the results. The publications in this project year are listed in the following.

Peer-Reviewd Journal Article:

1. Wei J, Chan HP, Helvie MA, Roubidoux MA, Sahiner B, Hadjiiski L, Zhou C, Paquerault S, Chenevert T, Goodsitt MM. Correlation between Mammographic Density and Volumetric Fibroglandular Tissue Estimated on Breast MR Images. Medical Physics 2004; 31: 933-942.

Articles Accepted for Publication:

1. Hadjiiski L, Chan HP, Sahiner B, Helvie MA, Roubidoux MA, Blane C, Paramagul C, Petrick N, Bailey J, Klein K, Foster M, Patterson S, Adler D, Nees A, Shen J. Improvement of radiologists' characterization of malignant and benign breast masses in serial mammograms by computer-aided diagnosis: An ROC study. Radiology.

Conference Proceedings:

1. Chan HP, Sahiner B, Hadjiiski LM, Petrick N, Zhou C. Design of three-class classifiers in computer-aided diagnosis: Monte Carlo simulation study. Proc SPIE 5032; 2003: 567-578.

2. Hadjiiski L, Chan HP, Sahiner B, Helvie MA, Roubidoux MA, Blane C, Paramagul C, Petrick N, Bailey J, Klein K, Foster M, Patterson S, Adler D, Nees A, Shen J. ROC study: Effects of computer-aided diagnosis on radiologists' characterization of malignant and benign breast masses in temporal pairs of mammograms. Proc SPIE 5032; 2003: 94-101.

3. Zhou C, Chan HP, Sahiner B, Hadjiiski LM, Paramagul C. Computerized multiple image analysis on mammograms: performance improvement of automated nipple identification for registration of multiple views using texture and geometric convergence analyses. Proc SPIE 5370; 2004, (in press).

4. Wei J, Sahiner B, Hadjiiski LM, Chan HP, Petrick N, Helvie MA, Zhou C, Ge Z. Computer aided detection of breast masses on full-field digital mammograms: false positive reduction using gradient field analysis. Proc SPIE 5370; 2004, (in press).

5. Hadjiiski LM, Helvie MA, Sahiner B, Chan HP, Roubidoux MA, Nees A, Petrick N, Blane C, Paramagul C, Bailey J, Patterson S, Klein K, Adler D, Foster M, Shen J. ROC Study of the Effects of Computer-Aided Interval Change Analysis on Radiologists' Characterization of Breast Masses in Two-View Serial Mammograms. Proc SPIE 5370; 2004, (in press).

6. Sahiner B, Chan HP, Hadjiiski LM, Roubidoux MA, Paramagul C, Helvie MA, Zhou C. Multi-modality CAD: Combination of computerized classification techniques based on mammograms and 3D ultrasound volumes for improved accuracy in breast mass characterization. Proc SPIE 5370; 2004, (in press).

Conference Abstracts and Presentations:

1. Chan HP, Sahiner B, Hadjiiski LM, Zhou C, Petrick N, "Three-Class Classification Tasks in Computer-Aided Diagnosis," Presented at *Medical Image Perception Society Conference,* Durham, NC, September 11-14, 2003.

2. Sahiner B, Chan HP, Hadjiiski LM, Zhou C, Wei J, Petrick N, "Comparison of resampling techniques for classifier performance estimation: the effects of feature selection, feature space dimensionality, and design sample size," Presented at *Medical Image Perception Society Conference,* Durham, NC, September 11-14, 2003.

3. Petrick N, Hadjiiski LM , Chan HP, Sahiner B, Myers KJ, "Assessment of a mammographic mass detection scheme with clinical cases," Presented at *Medical Image Perception Society Conference,* Durham, NC, September 11-14 2003.

4. Sahiner B, Chan HP, Hadjiiski LM, Helvie MA, Roubidoux MA, Petrick N. Computerized detection of microcalcifications on mammograms: Improved detection accuracy by combining features extracted from two mammographic views. Presented at the 89th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 30-December 5, 2003. RSNA Program 2003; 389.

5. Chan HP, Wei, J, Zhou C, Helvie MA, Roubidoux MA, Bailey J, Paramagul C, Nees A, Hadjiiski LM, Sahiner B. Comparison of mammographic density estimated on digital mammograms and screen-film mammograms. Presentation at the 89th Scientific

Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 30-December 5, 2003. RSNA Program 2003; 424.

6. Wei, J, Sahiner B, Chan HP, Petrick N, Hadjiiski LM, Helvie MA. Computer-aided diagnosis system for mass detection: Comparison of performance on full-field digital mammograms and digitized film mammograms. Presented at the 89th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 30-December 5, 2003. RSNA Program 2003; 387.
7. Hadjiiski LM, Chan HP, Sahiner B, Zhou C, Helvie MA, Roubidoux MA. Computerized Regional Registration of Corresponding Masses and Microcalcification Clusters on Temporal Pairs of Mammograms for Interval Change Analysis. Presented at the 89th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 30-December 5, 2003. RSNA Program 2003; 389.
8. Sahiner B, Chan HP, Roubidoux MA, Helvie MA, Bailey J, Hadjiiski LM. An ROC Study on Characterization of Malignant and Benign Breast Masses in 3D Ultrasound Volumes: The Effect of Computer-Aided Diagnosis on Radiologists' Characterization Accuracy. Presented at the 89th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 30-December 5, 2003. RSNA Program 2003; 425.
9. Zhou C, Hadjiiski LM, Sahiner B, Chan HP, Helvie MA, Wei, J. Computerized mammographic breast density estimation: Expectation-Maximization estimation and neural network classification of breast density. Presented at the 89th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 30-December 5, 2003. RSNA Program 2003; 389.
10. Wei J, Sahiner B, Hadjiiski LM, Chan HP, Petrick N, Helvie MA, Zhou C, Ge Z. Computer aided detection of breast masses on full-field digital mammograms: false positive reduction using gradient field analysis. Poster presentation at the SPIE International Symposium on Medical Imaging, San Diego, CA, February, 2004.
11. Hadjiiski LM, Helvie MA, Sahiner B, Chan HP, Roubidoux MA, Nees A, Petrick N, Blane C, Paramagul C, Bailey J, Patterson S, Klein K, Adler D, Foster M, Shen J. ROC Study of the Effects of Computer-Aided Interval Change Analysis on Radiologists' Characterization of Breast Masses in Two-View Serial Mammograms. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February, 2004.
12. Zhou C, Chan HP, Sahiner B, Hadjiiski LM, Paramagul C. Computerized multiple image analysis on mammograms: performance improvement of automated nipple identification for registration of multiple views using texture and geometric convergence analyses. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February, 2004.
13. Sahiner B, Chan HP, Hadjiiski LM, Roubidoux MA, Paramagul C, Helvie MA, Zhou C. Multi-modality CAD: Combination of computerized classification techniques based on mammograms and 3D ultrasound volumes for improved accuracy in breast mass characterization. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February, 2004.

14. Chan HP, Wei J, Sahiner B, Hadjiiski LM, Helvie MA, Petrick N, Roubidoux MA, Zhou C. Computer-aided diagnosis system of detection of breast masses on digital mammograms. Presented at the Peer Reviewed Medical Research Program (PRMRP) Investigators Meeting. Puerto Rico, April 26-28, 2004. Program book p.29.

(8) Conclusions

Under the support of this grant, we have investigated various computer-aided diagnosis (CAD) methods for detection of lesions on mammograms. We continue to collect a database of full field digital mammograms that contain mammographic lesions from our breast imaging division in the Department of Radiology. The images include the manufacturer's processed images and unprocessed (raw) images. All collected cases are entered into our database management program that stores the coded case information to facilitate archiving and retrieval of the cases. In this project year, we also extend our studies to improve the CAD system for DFMs. Since conventional film mammograms are more commonly available in patient files, we believe that it will be more efficient to first develop new computer-vision techniques and multiple-view image information fusion methods with a large number of DFMs, and then adapt the new methods to the CAD system for DMs. This approach has the additional advantages that the CAD system for DFMs will also be improved with the multiple-view information fusion method. Since DFMs will continue to be used for screening mammography in most breast imaging clinics in the near future, the improvement of the CAD systems for DFMs will benefit a large number of patients who will undergo conventional screen-film mammography. Although the development of CAD methods for DFMs was not included in our original proposed Statement of Work, this extension will strengthen our research and broaden its scope of applications to breast cancer diagnosis.

To facilitate the adaptation of the computer-vision techniques and the CAD system for DFMs to DMs, we investigated the image properties of DFMs and DMs taken of the same breasts within a short period of time. We compared the mammographic density estimated from the two types of mammograms by experienced radiologists. It was found that the correlation of the segmented breast density between the two types of images is very high. However, the estimated percent dense area on digital mammograms was, on average, about 3.5% lower than that estimated from digitized film mammograms. The average ratio of percent dense area on DFMs to that on DMs was 1.3. These results indicate that the perceived density on DMs is lower than that on DFMs. This difference in mammographic density may lead to improved sensitivity for lesion detection on digital mammograms both by radiologists and by the computer.

We are developing two-view information fusion method for correlating lesion detection and characterization on the CC and MLO views of mammograms, similar to radiologists' approach for mammographic interpretation. A regional registration technique is being developed to correlate the locations of the same lesion on different views. In this method, the nipple is used as the landmark for alignment of the breast images from the two views in a polar coordinate system. In this project year, we investigated methods to automatically detect the nipple location on a mammogram. We found that our current

method can identify about 85% of the nipples within 1 cm of the true location marked by radiologists. This result is promising although further improvement is still needed.

We are also developing computer-vision methods for classification of malignant and benign masses. Both single-view classification and two-view classification are being studied. In this project year, we performed a study to compare the classification accuracy with single-view and fused two-view information and the effects of CAD on radiologists' characterization of masses. We found that the two-view fusion can improve the classifier's accuracy, and that CAD can significantly improve radiologists' accuracy in characterization of malignant and benign masses. These results showed that CAD may be useful for reducing unnecessary biopsies.

In summary, we have investigated a number of areas in CAD of mammographic lesions. We have made progress in the six tasks proposed in the project. This lays the strong foundation for us to continue the development of the CAD systems for digital mammograms and digitized film mammograms in the coming years.

(9) References

1. Dorfman DD, Berbaum KS, Metz CE. Roc rating analysis: Generalization to the population of readers and cases with the jackknife method. Investigative Radiology 1992; 27: 723-731.

(10) Appendix

Copies of the following publications are enclosed with this report.

Peer-Reviewd Journal Article:

1. Wei J, Chan HP, Helvie MA, Roubidoux MA, Sahiner B, Hadjiiski L, Zhou C, Paquerault S, Chenevert T, Goodsitt MM. Correlation between Mammographic Density and Volumetric Fibroglandular Tissue Estimated on Breast MR Images. Medical Physics 2004; 31: 933-942.

Conference Proceedings:

1. Chan HP, Sahiner B, Hadjiiski LM, Petrick N, Zhou C. Design of three-class classifiers in computer-aided diagnosis: Monte Carlo simulation study. Proc SPIE 5032; 2003: 567-578.

2. Hadjiiski L, Chan HP, Sahiner B, Helvie MA, Roubidoux MA, Blane C, Paramagul C, Petrick N, Bailey J, Klein K, Foster M, Patterson S, Adler D, Nees A, Shen J. ROC study: Effects of computer-aided diagnosis on radiologists' characterization of malignant and benign breast masses in temporal pairs of mammograms. Proc SPIE 5032; 2003: 94-101.

Correlation between mammographic density and volumetric fibroglandular tissue estimated on breast MR images

Jun Wei,[a] Heang-Ping Chan, Mark A. Helvie, Marilyn A. Roubidoux, Berkman Sahiner, Lubomir M. Hadjiiski, Chuan Zhou, Sophie Paquerault, Thomas Chenevert, and Mitchell M. Goodsitt
Department of Radiology, University of Michigan, Ann Arbor, Ann Arbor, Michigan 49109

(Received 18 June 2003; revised 26 November 2003; accepted for publication 21 January 2004; published 26 March 2004)

Previous studies have found that mammographic breast density is highly correlated with breast cancer risk. Therefore, mammographic breast density may be considered as an important risk factor in studies of breast cancer treatments. In this paper, we evaluated the accuracy of using mammograms for estimating breast density by analyzing the correlation between the percent mammographic dense area and the percent glandular tissue volume as estimated from MR images. A dataset of 67 cases having MR images (coronal 3-D SPGR T1-weighted pre-contrast) and corresponding 4-view mammograms was used in this study. Mammographic breast density was estimated by an experienced radiologist and an automated image analysis tool, Mammography Density ESTimator (MDEST) developed previously in our laboratory. For the estimation of the percent volume of fibroglandular tissue in breast MR images, a semiautomatic method was developed to segment the fibroglandular tissue from each slice. The tissue volume was calculated by integration over all slices containing the breast. Interobserver variation was measured for 3 different readers. It was found that the correlation between every two of the three readers for segmentation of MR volumetric fibroglandular tissue was 0.99. The correlations between the percent volumetric fibroglandular tissue on MR images and the percent dense area of the CC and MLO views segmented by an experienced radiologist were both 0.91. The correlation between the percent volumetric fibroglandular tissue on MR images and the percent dense area of the CC and MLO views segmented by MDEST was 0.91 and 0.89, respectively. The root-mean-square (rms) residual ranged from 5.4% to 6.3%. The mean bias ranged from 3% to 6%. The high correlation indicates that changes in mammographic density may be a useful indicator of changes in fibroglandular tissue volume in the breast. © *2004 American Association of Physicists in Medicine.* [DOI: 10.1118/1.1668512]

Key words: mammography, breast density, MR images, correlation

I. INTRODUCTION

Studies have shown that there is a strong positive correlation between breast parenchymal density imaged on mammograms and breast cancer risk.[1–3] The relative risk is estimated to be about 4 to 6 for women whose mammograms have parenchymal densities over 60% of the breast area, as compared to women with less than 5% densities. Other cohort studies[4–13] also found that breast cancer risk in the category with the most extensive dense tissue was 1.8 to 6 times as high as that in the category with the least extensive dense tissue. Mammographic density as the risk indicator is greater than almost all other risk factors of breast cancer.[2,14] Although there is no direct evidence that changes in mammographic breast densities will result in changes in breast cancer risk, the strong correlation between breast density and breast cancer risk has prompted researchers to use mammographic density for monitoring the effects of intervention as well as for studying breast cancer etiology.[14–17]

A number of researchers have investigated image analysis techniques to estimate breast density.[15,18–28] The common approaches are to analyze the textural pattern or the percentage of mammographic densities relative to the breast area. It has been found that the texture measures were correlated with parenchymal density patterns but they appeared to be less sensitive measures of relative risk than the percent dense area.[1,25,29] In current practice, breast density is estimated mainly by radiologists' visual judgment of the fibroglandular tissue imaged on mammograms following the Breast Imaging—Reporting and Data System (BI-RADS) lexicon.[30,31] Because of the qualitative and subjective nature of visual judgment, there are large intraobserver and interobserver variations in the estimated breast density. The large variability may reduce the observed correlation between breast cancer risk and breast density. It may also reduce the sensitivity of studies using mammographic density for monitoring the effect of risk modifying treatments. We have developed an automated image analysis system, Mammographic Density ESTimator (MDEST), to assist radiologists in estimating breast density on mammograms. A computerized analysis is expected to increase the reproducibility and consistency in the estimation of mammographic density, thereby improving the accuracy of the related studies. In our previous study, we have found that the percent mammographic density segmented by MDEST agreed closely with that estimated by radiologists' interactive thresholding.[32]

The high correlation between breast cancer risk and breast

density indicates that breast cancer risk may be closely related to the volume of glandular tissue in the breast. Among the modalities available for breast imaging at present, magnetic resonance (MR) imaging is likely to be the most accurate method for volumetric dense tissue estimation because fibroglandular tissue and adipose tissue can be well distinguished in MR images when a proper image acquisition technique is used.[33] However, MR imaging is expensive, making it difficult to use MR imaging as a routine monitoring tool.[33,34] On the other hand, a mammogram is a two-dimensional (2-D) projection image of a three-dimensional (3-D) object. The area of dense tissue measured on a mammogram is not an accurate measure of the volume of fibroglandular tissue in the breast because no thickness information is used. However, mammography is a widely available low cost procedure that may be used for monitoring breast density change during preventive and interventional treatment or other studies. Women who participate in screening will also have mammograms readily available for retrospective review. Therefore, mammography will most likely be the method of choice for breast density estimation.

In this study, we investigated the correlation between the volumetric fibroglandular tissue in the breast and the projected breast dense area on mammograms by analyzing the percent volumetric fibroglandular tissue in MR breast images and the percent dense area in corresponding mammograms. Our purpose in this study is not to evaluate the usefulness of either MR fibroglandular tissue volume or mammographic density as an indicator for breast cancer risk, which have been studied by other investigators. Rather, we used the MR breast images to estimate the volumetric fibroglandular tissue in the breast and explored the reason that a change in mammographic density (2-D) can be used as an indicator of breast density change (3-D). These comparisons will provide a better understanding of their relationship, and may lead to improved methods for utilizing mammographic density as a surrogate marker for breast cancer risk.

II. MATERIALS AND METHOD

A. Dataset

In a previous study, gadolinium contrast enhanced MR dynamic imaging was employed to characterize malignant and benign breast lesions. A dataset was collected with IRB approval which included MR images and corresponding mammograms acquired between detection and before biopsy for a given patient. In the MR study, several series of images were acquired for each patient. Patients were scanned prone using a commercial dual phased-array breast coil. The imaging protocol included a series was the coronal 3-D T1-weighted pre-contrast series (coronal sections 2–5 mm thick, 32 slices; 3-D Spoiled Gradient-Recalled Echo (SPGR); TE $=3.3$ ms; TR$=10$ ms, Flip$=40°$, matrix$=256\times128$, FOV $=28$–32 cm right/left, 14–16 cm superior/inferior, scan time$=2$ min 38 sec). This 3-D SPGR sequence produces full volume coverage of both breasts with contiguous image sections. The dense parenchyma and fat tissue are well separated with this heavily T1-weighted acquisition. We used a set of 67 patients to study the correlation between the 2-D projected percentage of dense area on a mammogram and the percentage of dense tissue volume estimated from the 3-D MR images.

The mammograms consisting of the craniocaudal (CC) view and the mediolateral oblique (MLO) view of both breasts of the patient were digitized with a LUMISYS 85 laser film scanner at a pixel size of $50\ \mu m\times50\ \mu m$. The digitizer has a gray level resolution of 12 bits and a nominal optical density (O.D.) range of 0 to 4. For density segmentation, it is not necessary to use very high-resolution images. To reduce processing time, the full resolution mammograms were first smoothed with a 16×16 box filter and subsampled by a factor of 16, resulting in $800\ \mu m\times800\ \mu m$ images for this study.

B. Estimation of fibroglandular tissue volume on MR images

Since it is not our intention to routinely segment MR images for breast density estimation, we did not attempt to develop an automated method for this application. Our algorithm for segmentation of volumetric fibroglandular tissue on MR images used a semi-automatic method. The computer performed an initial segmentation. A graphical user interface (GUI) was developed to allow a user to review the segmentation of every slice and make modifications if necessary. The method consists of four steps. First, the breast boundary was detected automatically on each slice. A deformable model and manual modification were used to correct for incorrectly detected boundaries that usually occurred in slices near the chest wall where there were no well-defined breast boundaries. Because of inhomogeneity of the breast coil sensitivity, the signal intensity in the breast region was not uniform across the field of view. A background correction technique that estimated the low frequency background from the gray levels along the breast boundary was developed to reduce this systematic nonuniformity. Manual interactive thresholding of the gray level histogram in the breast region was then used to separate the fibroglandular from the fatty region. Morphological erosion was used to exclude the skin voxels along the breast boundary. Finally, the volume of fibroglandular tissue was calculated by integration over all slices containing the breast. A flow chart of our algorithm is shown in Fig. 1.

C. Breast boundary detection

A two-step algorithm was developed for the detection of breast boundary on each slice. First, we used a seeded pixel thresholding algorithm (SPTA) for the initial assessment of a breast boundary. Second, a 2-D active contour algorithm further refined the boundary. For slices close to the chest wall where no clear boundary can be seen, manual modification was used to outline an estimated boundary.

The SPTA determined the optimal threshold by iteratively partitioning the MR image into two parts and using the gradient value along the boundary of the partition as a guide in optimizing the threshold. First, the center of gravity was se-

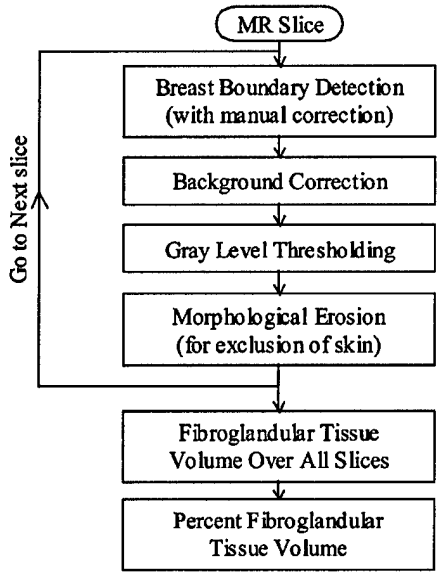


FIG. 1. The flow-chart for the segmentation of the fibroglandular tissue on MR images.

lected as the starting pixel on each slice. The gray level of the starting pixel was used as a threshold to create a binary partition of the image in which all pixels greater than the threshold were set to one and all other pixels were set to zero. Second, the gradient value of each pixel on the boundary of the binary partition was calculated by applying the Sobel filter to the original image. The gradient assessment for this particular binary partition was defined as the average gradient magnitude of these boundary pixels. The threshold value was reduced to zero in a stepwise manner. The partition for each threshold value was created and the gradient assessment for each partition was calculated as described above. The partition with the maximum gradient assessment was considered to be the initial segmentation result for the breast, and the boundary of this partition was considered to be the initial breast boundary.

After the initial segmentation, a deformable contour method was used to further refine the boundary. The movement of the boundary pixel was controlled by an energy function which consisted of internal energy and external energy. The internal energy components used in this study were the continuity and curvature of the contour, as well as the homogeneity of the segmented partition. The external energy components were the negative of the smoothed image gradient magnitude, and a balloon force that exerted pressure at a normal direction to the contour. The energy function was defined as the following:

$$E=\sum_{c=1}^{N}[E_{\text{inter}}(c)+E_{\text{exert}}(c)], \tag{1}$$

where E_{inter} and E_{exert} are the internal energy and the external energy, respectively, as defined in Eq. (2) and Eq. (3):

$$E_{\text{inter}}=w_{\text{curv}}E_{\text{curv}}(c)+w_{\text{cont}}E_{\text{cont}}(c)+w_{\text{hom}}E_{\text{hom}}, \tag{2}$$

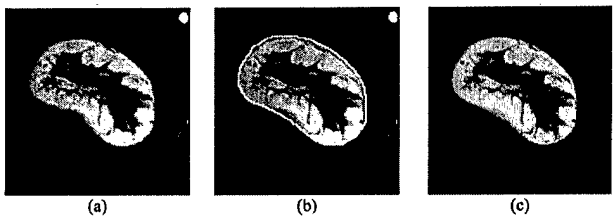


FIG. 2. An example of the first three processing blocks in Fig. 1. (a) Original MR slice; (b) automatically-detected breast boundary superimposed on the image; and (c) the background-corrected image.

$$E_{\text{exert}}=w_{\text{grad}}E_{\text{grad}}(c)+w_{\text{bal}}E_{\text{bal}}(c), \tag{3}$$

where curv, cont, grad, bal, hom denoted curvature, continuity, gradient, balloon force and homogeneity, respectively, and each energy term was associated with a weight, w. The detailed definition for each term can be found in the literature.[35] An example of a MR slice of a breast is shown in Fig. 2(a), and the segmented boundary is shown in Fig. 2(b). Note that the two breasts of a patient were scanned together but each breast was analyzed separately.

D. Background correction

To reduce the nonuniformity of the MR signal intensity in the breast region, a background correction technique[36] using the pixel values around the segmented breast region was employed. For a given pixel (i,j) inside the breast region, the gray value of the background image was estimated as shown in Eq. (4):

$$B(i,j)=\left[\frac{L}{d_l}+\frac{R}{d_r}+\frac{U}{d_u}+\frac{D}{d_d}\right]\Big/\left[\frac{1}{d_l}+\frac{1}{d_r}+\frac{1}{d_u}+\frac{1}{d_d}\right], \tag{4}$$

where L, R, U and D are the average gray values inside a breast background estimation region (BBER) centered at the left, right, upper and lower pixels on the breast boundary, respectively. A BBER was defined as the intersection of a 21×21-pixel box and the breast region. The center pixels for the left and right boxes were the intersection points between the breast boundary and a horizontal line passing through the given pixel (i,j). Similarly, the upper and lower center pixels for the upper and lower boxes were the intersection points between the breast boundary and a vertical line passing through the given pixel (i,j). Only the pixels that were within the intersected area between the 21×21-pixel box and the breast region were included in the definition of the BBER and the calculation of the average gray value. The contributions of the average gray levels to the background pixel (i,j) were inversely weighted by their distances d_l,d_r,d_u,d_d from the given pixel (i,j). An example of the background corrected image is shown in Fig. 2(c).

E. Segmentation of fibroglandular tissue

We developed a GUI that allowed the user to perform a combination of manual and automatic operations to segment the breast boundary and the fibroglandular tissue on the MR

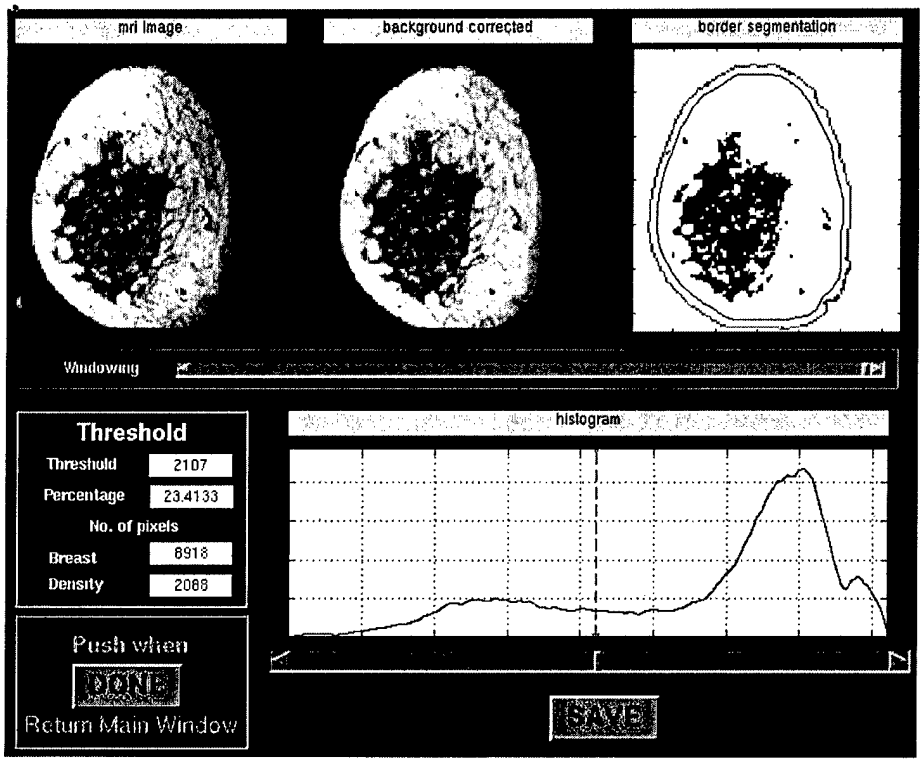


FIG. 3. The graphic user interface for the segmentation of the fibroglandular tissues on the MR slice. The upper row shows the original MR slice (left), the background-corrected image (middle) and the segmented binary image (right). The segmented image responds to the reader's adjustment of the gray level threshold (lower row) in real time so that the reader can choose the appropriate threshold by inspecting the segmented image visually. The dark area in the segmented image indicates the fibroglandular tissue and the white area indicates the adipose tissue. The inner line along the breast boundary is the boundary obtained by morphological erosion to exclude the skin voxels for calculating the fibroglandular tissue volume.

images. The first window (not shown) displayed the MR series and the corresponding mammogram of each breast to give the user an overview of the breast. The segmentation of the fibroglandular tissue on each MR slice was processed in the second window, shown in Fig. 3. The original MR slice, the corresponding background corrected image and the segmented binary image were shown in the upper part of the window. At the lower part of the window, the histogram of the voxel values in the breast region was shown. The user performed interactive thresholding on the histogram and the segmented binary image corresponding to the chosen threshold was displayed in real time in the upper part. If the breast boundary, which was automatically segmented by the computer initially, had to be corrected, the user could go to the third window and manually move the apices of the polygon outlining the boundary. The voxels contributed by the nipple were excluded. On the slices containing breast skin that had voxel values similar to those of fibroglandular tissue, a morphological erosion operation was applied to the breast boundary to exclude the skin voxels from the calculation of the fibroglandular tissue volume in the slice. The size of the structuring element could be selected interactively on the fourth window and the eroded boundary was displayed instantly for a chosen erosion operation. The user might again change the structuring element if the erosion result of the previous choice was deemed unsatisfactory. Since the eroded boundary only marked the region within which the fibroglandular voxels would be summed and would not be used for the calculation of the breast volume, as described below, it did not need to be precise as long as it excluded the skin voxels while not excluding the fibroglandular voxels.

F. MR fibroglandular tissue volume

After the fibroglandular tissue was segmented for each slice, the total number of voxels containing the fibroglandular tissue was obtained as a summation of these voxels over all slices of the breast. The total volume of the breast was obtained as the summation of the voxels enclosed by the breast boundary before morphological erosion. The ratio of these two volumes provided the percent volumetric fibroglandular tissue in the breast.

G. Mammographic density segmentation

We have previously developed an automated method for segmentation of the dense fibroglandular area on mammograms. The method, referred to as the Mammographic Density ESTimator (MDEST) was described in detail elsewhere.[32] In brief, the breast boundary on the digitized mammogram is tracked. A dynamic-range compression technique reduces the gray level range of the breast area. By analyzing the shape of the gray level histogram, a rule-based classifier classifies the breast density into one of four classes. Typically, a Class I breast is almost entirely fat; it has a single narrow peak on the histogram. A Class II breast contains scattered fibroglandular densities. Its histogram has two main peaks, with the smaller peak on the right of the bigger one. A Class III breast is heterogeneously dense. Its histogram also has two peaks, but the smaller peak is on the left of the bigger one. A Class IV breast is extremely dense. Its histogram has mainly a single dominant peak, but the peak is wider compared with the peak in the Class I histogram. A second smaller peak sometimes occurs on the left of the

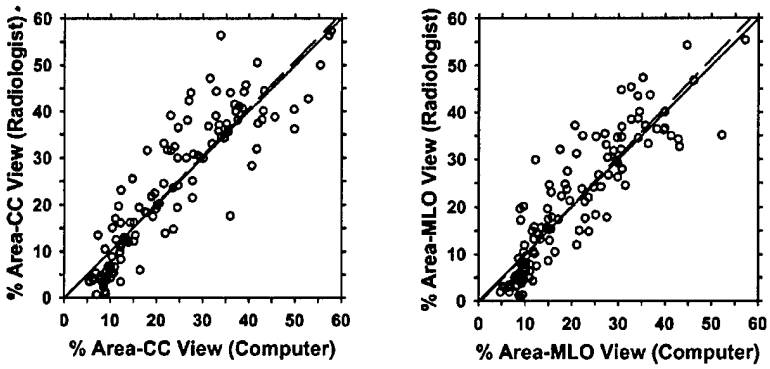


FIG. 4. A comparison of the percent mammographic density obtained from interactive thresholding by an MQSA-qualified radiologist and that estimated by our automated MDEST computer program. (a) CC view, correlation coefficient=0.90, rms residual=6.7, mean difference=0.3; (b) MLO view, correlation coefficient =0.89, rms residual=6.1, mean difference=0.4. Dashed line: linear regression of the data; solid line: diagonal.

main peak. Based on the histogram shape, a threshold is automatically calculated to separate the dense and fatty pixels. The mammographic density was estimated as the percentage of fibroglandular tissue area relative to the total breast area. For MLO view mammograms, the pectoral muscle is detected and excluded from the density area or breast area calculations. In our previous work, the performance of MDEST was verified by comparison with manual segmentation by 5 breast imaging radiologists using a dataset of 260 mammograms from 65 patients that were different from the cases used in the current study. We found that the correlation between the computer-estimated percent dense area and the average segmentation by the 5 radiologists was 0.94 and 0.91, respectively, for CC and MLO views, with a mean bias of less than 2%.

MDEST was applied to the mammograms of the 67 patients used in this study. The percent dense area on mammograms was estimated for the CC-view and the MLO-view mammogram of each breast separately. In addition, an MQSA-qualified radiologist also segmented the dense area by interactive thresholding for each mammogram. The correlation between the mammographic density obtained by manual and automatic segmentation is shown in Figs. 4(a) and 4(b) for the CC view and MLO view, respectively. The correlation coefficients for the CC view and MLO view were 0.90 and 0.89, respectively. The mammographic densities estimated by automatic and manual segmentation were compared with the percent volumetric fibroglandular tissue on MR images as described below.

H. Observer experiments

We performed an experiment to evaluate the variability of the estimated % volumetric fibroglandular tissue due to the uncertainty in the determination of the starting slice of the breast at the chest wall. The starting slice affected the estimation of the breast volume that was calculated by integrating from the starting slice to the anterior of the breast. Twenty-three MR cases from the dataset were randomly selected for this observer experiment. There were a total of 41 breasts because some cases had only one breast. For this subset of cases, each radiologist was asked to select the starting slice from the MR images for each breast. The estimated % volumetric fibroglandular tissue calculated with all available slices was then compared to that calculated with the selected starting slice.

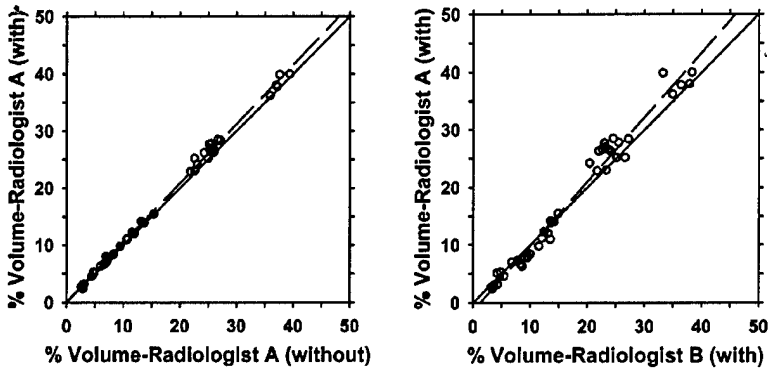
We also performed observer experiments to evaluate the inter-observer variations in the segmentation of fibroglandular tissue using the semi-automatic method. Two MQSA-qualified radiologists performed the segmentation of the fibroglandular tissue on the MR images of the 41 breasts using the semi-automatic method implemented with the GUI. A Ph.D. researcher who was trained by these radiologists also performed the segmentation independently with the GUI.

After verifying the consistency of segmentation by these observers, the trained Ph.D. completed the segmentation of all MR cases. The correlation between percent volumetric fibroglandular tissue on MR images and percent dense area on mammograms was then examined for the entire dataset.

III. RESULTS

A. Effect of selection of the starting slice

Figure 5(a) shows the correlation of the % volumetric fibroglandular tissue calculated using all available slices for the breast with that calculated using the selected starting slice by radiologist A for the 41 breasts. The correlation coefficient was 0.999. To compare the difference between their results, the mean difference and the root-mean-square (rms) residual, which is the residual from the linear least-squares-fitted line, were also calculated. The mean difference was 0.7 and the rms residual was 0.6. The result is similar for radiologist B (not shown), with a correlation coefficient of 0.999, a mean difference of 0.4 and a rms residual of 0.4. The correlation between the % volumetric fibroglandular tissue calculated using the selected starting slice by radiologist A with that calculated using the selected starting slice by radiologist B was also very high with a correlation coefficient of 0.988, a mean difference of 0.7 and a rms residual of 1.8, as shown in Fig. 5(b). These comparisons indicated that the variability in the selection of the starting slice of the breasts did not have a strong influence on the % volumetric fibroglandular tissue. We therefore used all available slices in the MR dataset for each breast in the following analyses.



(a) (b)

FIG. 5. (a) A comparison of the percent fibroglandular tissue volume calculated using the selected starting slice with that calculated using all available slices for radiologist A, correlation coefficient=0.999. (b) A comparison of the percent fibroglandular tissue volume calculated using the selected starting slice by radiologist B with that by radiologist A, correlation coefficient =0.988, Dashed line: linear regression of the data; solid line: diagonal.

B. Inter-observer variation between radiologists

Figure 6(a) shows the comparison of the percent volumetric fibroglandular tissues on MR images segmented by two radiologists for the 41 breasts. The correlation between the segmentation results of the two radiologists is 0.99. The mean difference was found to be 0.3 and the rms residual was 1.6.

C. Inter-observer variation between radiologists and trained Ph.D.

Figure 6(b) shows the comparison of the percent volumetric fibroglandular tissues segmented by the trained Ph.D. against that segmented by radiologist A. A similar result was obtained by comparing the percent volumetric tissue segmented by the trained Ph.D. and that segmented by radiologist A except that the data points were even closer to the diagonal (not shown). The correlation between the result of the trained Ph.D. and the results of both radiologists was 0.99. The corresponding mean differences were -0.8 and -0.4, respectively, and the rms residuals were 1.4 and 1.5, respectively.

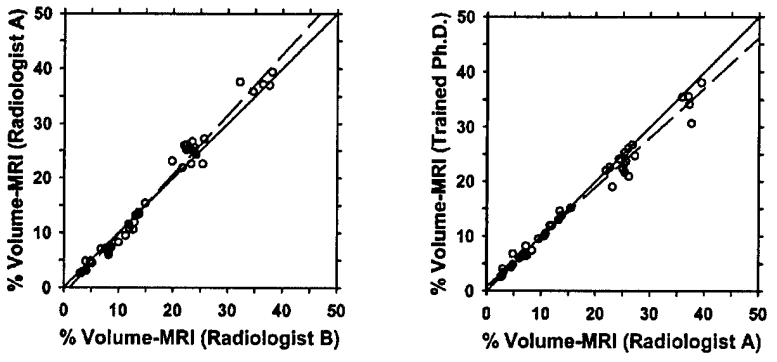
D. Correlation between percent volumetric fibroglandular tissue on MR images and percent mammographic density

The percent volumetric fibroglandular tissue on MR images was compared with the percent dense area on CC- and MLO-view mammograms. After verifying that the difference in segmentation between the trained Ph.D. and the radiologists was similar to the interobserver variations between the two experienced radiologists, the trained Ph.D. completed the segmentation of the entire dataset.

Figure 7 shows the comparison of the percent volumetric fibroglandular tissue on MRI and the percent mammographic density segmented by a radiologist. The percent areas on CC- and MLO-view mammograms are higher than the percent volume on MR images with a mean difference of 5.7% and 3.0%, respectively.

Figure 8 shows the comparison of the percent volumetric fibroglandular tissue on MRI and the percent mammographic density segmented by MDEST. The percent areas on CC- and MLO-view mammograms segmented by the computer are higher than the percent volume on MR images with a mean difference of 5.3% and 2.6%, respectively.

The correlation coefficients, the mean differences and the rms residuals between the percent volumetric fibroglandular tissue on MR images and percent dense area on mammograms are compared in Table. I. The correlation between the percent volume on MR images and percent area on mammograms of the fibroglandular breast tissue is high, ranging from 0.89 to 0.91. Although it is not expected that the values of percent volume agree with the values of percent area, their mean differences range only from 3% to 6% and the rms residual range from 5.4 to 6.3.



(a) (b)

FIG. 6. A comparison of the segmentation of fibroglandular tissue from MR images between two observers: (a) two experienced MQSA-qualified radiologists, correlation coefficient=0.99. (b) The trained Ph.D. and Radiologist A, correlation coefficient=0.99. The correlation between the trained Ph.D. and Radiologist B is also 0.99 but the data points were very close to the diagonal and is not shown. The % volumetric fibroglandular tissue was calculated using all available slices. Dashed line: linear regression of the data; solid line: diagonal.

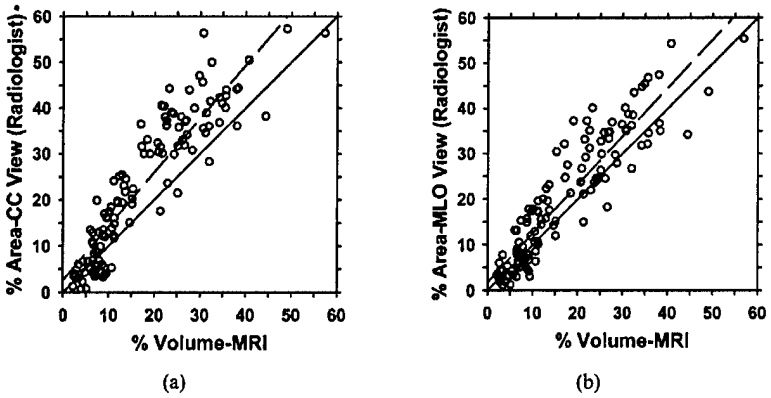


FIG. 7. A comparison of the percent fibroglandular tissue volume on MR images and the percent dense area on mammograms segmented by an experienced radiologist. (a) CC view, correlation coefficient=0.91; (b) MLO view, correlation coefficient=0.91. Dashed line: linear regression of the data; solid line: diagonal.

IV. DISCUSSION

Our purpose in this paper was to investigate the relationship between the percent dense area on mammogram and the percent fibroglandular tissue volume on MR image. We found a direct correlation between mammographic density and MR volumetric density (Fig. 7 and Fig. 8). The correlation coefficients between the percent area on a mammogram and the percent volume on MR images are high at 0.89 and 0.91. These results are more promising than those found in previous studies that attempted to correlate percent dense area on mammograms with MR information. Graham *et al.*[33] investigated the relationship between percent density (projected dense area) on mammogram and two objective MR parameters of breast tissue, relative water content and mean T2 relaxation. Their results with 45 cases showed a positive correlation between percent density and relative water content (Pearson correlation coefficient=0.79) and a negative correlation between percent density and mean T2 value (Pearson correlation coefficient=−0.61). Another study by Lee *et al.*[34] analyzed fatty and fibroglandular tissue in different age groups to compare x-ray mammography with T1-weighted MR images. Their study with 40 cases indicated that the correlation between the two techniques is 0.63 when the fat content was more than 45%. However, the correlation coefficient decreased to 0.34 when their analysis included only dense breasts.

It may be noted that although MR imaging is currently the most accurate method for estimating the volumetric fibroglandular tissue in the breast, it is still not the ideal tool. Fibrous tissue and glandular tissue are not well separated with current MR imaging techniques. Since the amount of glandular tissue in the breast is the important factor relating to breast cancer risk, further studies are warranted for differentiating the glandular and the fibrous components of the imaged volume. The correlation between the percent glandular tissue volume and percent projected dense area on a mammogram will be a more reliable indicator of the usefulness of mammographic density analysis.

The density on mammograms is a 2-D projected area of the fibroglandular tissues. The percent dense area is not expected to be equal in value to the percent volume. The mean differences between the percent volume and the percent area on CC- and MLO-views, as determined by the radiologist's interactive segmentation, are 5.7 and 3.0, respectively (Table I), with the percent dense area values being higher. We also investigated the rms residual between the percent volume and the percent area when the relationship between them was assumed to be linear. The rms residual between the percent volume and the percent area on CC- and MLO-views are 6.3 and 5.6, respectively (Table I), relative to the straight line obtained from linear least squares fits to the data. One possible factor that may contribute to a higher value of percent dense area on mammograms than the percent volume value on MR images is that the tissue volume imaged by the two modalities is somewhat different. The MR images include more tissue near the chest wall, which is mainly retroglan-

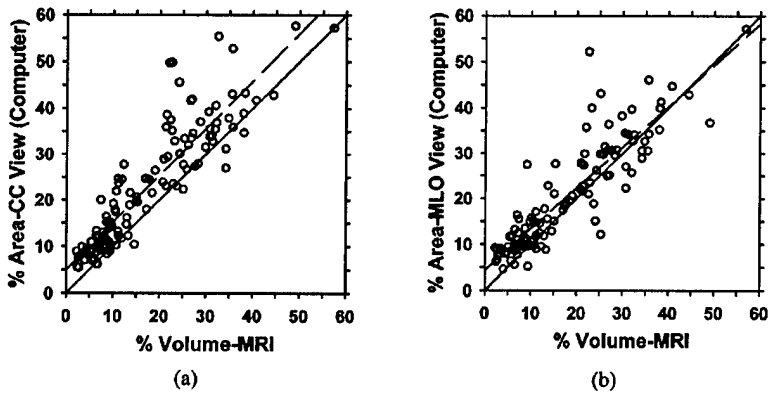


FIG. 8. A comparison of the percent volume on MR images and the percent area on mammogram segmented by our automated MDEST computer program. (a) CC view, correlation coefficient=0.91; (b) MLO view, correlation coefficient=0.89. Dashed line: linear regression of the data; solid line: diagonal.

TABLE I. Statistic analysis of the relationship between percent fibroglandular tissue volume on breast MR images and percent dense area on mammograms segmented by radiologist and MDEST.

	Radiologist		*Computer (MDEST)*	
	CC vs MRI	MLO vs MRI	CC vs MRI	MLO vs MRI
Correl. coeff.	0.91	0.91	0.91	0.89
rms residual	6.3	5.6	5.8	5.4
Mean diff.	5.7	3.0	5.3	2.6

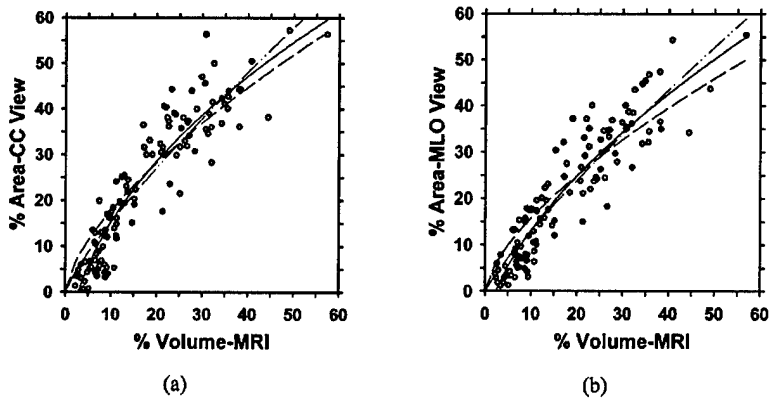
TABLE II. An analysis of the relationship between percent fibroglandular tissue volume (x) on breast MR images and percent dense area (Y) on mammograms segmented by radiologist using three mathematical models. m, k, k_1 and k_2 are constants determined by least squares curve fitting.

Mathematical model		$Y=kx^{2/3}$	$Y=kx^m$	$Y=k_1x^m+k_2$
CC vs MRI	Least squares Fit	$Y=0.82x^{2/3}$	$Y=1.03x^{0.83}$	$Y=1.02x^{0.48}-0.19$
	rms residual	6.5	6.0	5.6
	Coefficient of determination	0.82	0.85	0.87
MLO vs MRI	Least squares Fit	$Y=0.73x^{2/3}$	$Y=0.96x^{0.86}$	$Y=0.90x^{0.60}-0.09$
	rms residual	6.0	5.5	5.3
	Coefficient of determination	0.80	0.84	0.85

dular adipose tissue, than a mammogram does, thus reducing the percentage of fibroglandular tissue volume. The reduction in the percent volume values, however, is relatively small, as found in our study evaluating the effects of selecting starting slices for volume calculation (Fig. 5). The main difference may therefore be attributed to the geometric relationship between the volume and the projected 2-D area, explained later.

Geometrically, we do not expect the relationship between volume and its projected 2-D area to be linear. In a hypothetical situation such that the dense tissue volume is a sphere (volume$=4/3\ \pi r^3$) enclosed inside a concentric spherical shell of fatty tissue volume, the percent projected 2-D area (area$=\pi r^2$) of the inner sphere relative to the outer sphere is equal to the percent volume to the power of 2/3. The relationship between the percent area and the percent volume is therefore not linear, and the percent area is larger in value than the percent volume for any ratio of radii between the two spheres. In general, the compressed breast and the dense tissue are not spherical. To investigate the empirical relationship between the percent area and the percent volume in the nonlinear situation, we applied least squares fits in several polynomial models to the data points in Fig. 7. The results are shown in Table II and Fig. 9. A comparison of Table I and Table II indicates that the $Y=kx^{2/3}$ model (x=percent fibroglandular tissue volume, Y=percent mammographic dense area) resulted in slightly larger rms residuals than the linear model. The model $Y=kx^m$ with m equal to 0.83 and 0.86, respectively, for CC- and MLO-views slightly reduced the rms residuals. The best fit was obtained from the model $Y=k_1x^m+k_2$. However, the situation that the percent projected area was negative when the percent volume was zero would not occur physically. Note that if the model was fitted to the percent area data segmented by MDEST (Fig. 8), the k_2 values would become positive, indicating that the nonzero k_2 values are likely caused by segmentation biases.

Overall, these models demonstrate that there is no simple mathematical relationship between the percent volume and the percent projected area but the values for the exponents appeared to be in a reasonable range. The relationship between the percent volumes of two 3-D objects, one within another, and their percent projected 2-D area depends on their shapes. For example, the closer the two volumes are to concentric cylinders of the same height, the closer the exponent is to unity. The spread of the data points can therefore be attributed to the various irregular shapes of the fibroglandular tissue in the breasts, the changes in the shapes of the fatty and fibroglandular tissue due to compression, as well as the uncertainties in the segmentation of both the mammograms and the MR images. Although the spread of the data points in the correlation plots is large, one can expect that when the mammographic density of a given patient is monitored over time, the variations in the projected dense area due to the geometric factors, described above, will actually be much less than that observed from the scatter plots among a large number of patients. In other words, the uncertainty in the estimated percent density from the serial mammograms of a given patient should be much less than those shown in



(a) (b)

FIG. 9. Nonlinear fitting of the relationship between the percent volume and the percent area segmented by a radiologist with the least squares method. (a) CC view, (b) MLO view. Dashed line: $y=kx^{2/3}$; dashed–dotted–dotted line: $y=kx^m$; solid line: $y=k_1x^m+k_2$. The fitted parameters of the models, m, k, k_1 and k_2, are shown in Table II.

Fig. 7. The strong correlation observed between the percent dense area on mammograms and the percent volumetric fibroglandular tissue on MR images therefore indicates that a change in mammographic density can be a useful indicator of a change in percent fibroglandular tissue volume in the breast.

Recently, some researchers attempted to estimate the thickness of the fibroglandular tissue in local regions of the mammograms from the projected density.[37] This approach is expected to provide a more accurate estimation of the fibroglandular tissue volume if the true thicknesses of the fibroglandular tissue and fatty tissue can be determined at various locations of the projected breast region. The volume of the fibroglandular tissue can then be summed over the pixels in the breast region and the percent volume calculated. However, to obtain accurate measurements, this approach requires the knowledge of the sensitometric curve for the screen-film mammogram at the imaging facility (or use of a digital detector with linear response) and other physical parameters such as the scatter fraction, the beam quality and beam hardening, in addition to the compressed breast thickness and the breast shape profile at the periphery. Some of the requirements may be circumvented by using a look-up table predetermined with a phantom calibration. Other factors may have to be approximated or ignored, or require further corrections by imaging each mammogram with a calibration phantom placed adjacent to the breast. This method is still being developed and the accuracy of estimating the thickness of the local fibroglandular tissue from a mammogram is yet to be determined. To our knowledge, no study to date has demonstrated that fibroglandular tissue volume estimated from mammograms has a higher correlation with the percent volumetric fibroglandular tissue volume estimated from MR images or other volumetric methods than we found in our current study. Furthermore, even if the local fibroglandular tissue thickness on mammograms can be measured in a laboratory or in an academic center using elaborate calibration schemes, it is doubtful that these methods can be translated into routine clinical measurement in mammography clinics. Its use may then be limited to controlled clinical trials. An estimation of the percent dense area projected on mammograms is likely a more practical approach for breast density assessment. The high correlation between the percent dense area and the percent fibroglandular tissue volume on MR images as demonstrated in the current study further supports the validity of this approach.

V. CONCLUSION

In this study, we investigated the correlation between the percent mammographic dense area and the percent volumetric fibroglandular tissue as measured on MR images. A semiautomatic method was developed for segmentation of the MR images and a fully automated computerized method, MDEST, was used to segment the mammograms. The performance of MDEST on the set of mammograms used in this study was verified with an experienced radiologist's manual segmentation. The inter-observer variability in segmentation of MR images was found to be small with correlation coefficients of 0.99. The correlation between the percent volume on MR images and percent area segmented by a radiologist for either CC- view or MLO-view is 0.91. The correlation between percent volume and percent area estimated by MDEST is 0.91 and 0.89, respectively, for CC and MLO views. Mammographic density is thus highly correlated with the percent volumetric fibroglandular tissue in the breast. The high correlation indicates that changes in mammographic density may be a useful indicator of changes in fibroglandular tissue volume in the breast. Our computerized image analysis tool, MDEST, can provide a consistent and reproducible estimation of percent dense area on routine clinical mammograms. The automated image analysis tool may improve the sensitivity of quantifying mammographic density changes, thereby contributing to the understanding of the relationship of mammographic density to breast cancer risk, detection, and prognosis, and the prevention and treatment of breast cancer.

ACKNOWLEDGMENTS

This work is supported by U.S. Army Medical Research and Materiel Command Grants No. DAMD 17-01-1-0326, No. DAMD 17-02-1-0214, and No. DAMD 17-99-1-9294. The content of this paper does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned should be inferred.

[a]Author to whom correspondence should be addressed. Jun Wei, Ph.D., Department of Radiology, University of Michigan, CGC B2103, 1500 E. Medical Center Drive, Ann Arbor, Michigan 48109. Phone: 734-647-8553; fax: 734-615-5513; electronic mail: jvwei@umich.edu

[1]A. F. Saftlas, R. N. Hoover, L. A. Brinton, M. Szklo, D. R. Olson, M. Salane, and J. N. Wolfe, "Mammographic densities and risk of breast cancer," Cancer (N.Y.) **67**, 2833–2838 (1991).

[2]N. F. Boyd, G. A. Lockwood, J. W. Byng, D. L. Tritchler, and M. J. Yaffe, "Mammographic densities and breast cancer risk," Cancer Epidemiology Biomarkers & Prevention **7**, 1133–1144 (1998).

[3]C. M. Vachon, C. C. Kuni, K. Anderson, V. E. Anderson, and T. A. Sellers, "Association of mammographically defined percent breast density with epidemiologic risk factors for breast cancer (United States)," Cancer Causes & Control **11**, 653–662 (2000).

[4]P. M. Krook, "Mammographic parenchymal patterns as risk indicators for incident cancer in a screening program: an extended analysis," Am. J. Roentgenolo. **131**, 1031–1035 (1978).

[5]R. L. Egan and R. C. Mosteller, "Breast cancer mammography patterns," Cancer (N.Y.) **40**, 2087–2090 (1977).

[6]B. Threatt, J. M. Norbeck, N. S. Ullman, R. Kummer, and P. Roselle, "Association between mammographic parenchymal pattern classification and incidence of breast cancer," Cancer (N.Y.) **45**, 2550–2556 (1980).

[7]M. Moskowitz, P. Gartside, and C. Mclaughlin, "Mammographic patterns as markers for high-risk benign breast disease and incident cancers," Radiology **134**, 293–295 (1980).

[8]I. Witt, H. S. Hansen, and S. Brunner, "The risk of developing breast cancer in relation to mammography findings," Eur. Radiol. **4**, 65–67 (1984).

[9]S. Ciatto and M. Zappa, "A prospective study of the value of mammographic pattern as indicators of breast cancer risk in a screening experience," Eur. Radiol. **17**, 122–125 (1993).

[10]E. Thurfjell, C. C. Hsieh, L. Lipworth, A. Ekbom, H. O. Adami, and D. Trichopoulos, "Breast size and mammographic pattern in relation to breast cancer risk," Eur. J. Cancer Prevention **5**, 37–41 (1996).

[11]I. Kato, C. Beinart, A. Bleich, S. Su, M. Kim, and P. G. Toniolo, "A nested case-control study of mammographic patterns, breast volume and

breast cancer (New York City, NY, United States)," Cancer Causes & Control **6**, 431–438 (1995).
[12] E. Sala, R. Warren, J. McCann, S. Duffy, N. Day, and R. Luben, "Mammographic parenchymal patterns and mode of detection: implications for the breast screening programme," J. Medical Screening **5**, 207–212 (1998).
[13] T. M. Salminen, I. E. Saarenmaa, M. M. Heikkila, and M. Hakama, "Is a dense mammographic parenchymal pattern a contraindication to hormonal replacement therapy?," Acta Oncol. **39**, 969–972 (2000).
[14] C. Byrne, C. Schairer, J. N. Wolfe, N. Parekh, M. Salane, L. A. Brinton, R. Hoover, and R. Haile, "Mammographic features and breast cancer risk: Effects with time, age, and menopause status," J. Natl. Cancer Inst. **87**, 1622–1629 (1995).
[15] N. F. Boyd, C. Greenberg, G. Lockwood, L. Little, L. Martin, J. Byng, Y. Martin, and D. Tritchler, "Effects at two years of a low-fat, high-carbohydrate diet on radiologic features of the breast: Results from a randomized trial," J. Natl. Cancer Inst. **89**, 466–467 (1997).
[16] D. V. Spicer, G. Ursin, Y. R. Parisky, J. G. Pearce, D. Shoupe, A. Pike, and M. C. Pike, "Changes in mammographic densities induced by a hormonal contraceptive designed to reduce breast cancer risk," J. Natl. Cancer Inst. **86**, 431–436 (1994).
[17] J. Brisson, R. Verreault, A. S. Morrison, D. Tennina, and F. Meyer, "Diet, mammographic features of breast tissue, and breast cancer risk," Am. J. Epidemiol. **130**, 14–24 (1989).
[18] J. N. Wolfe, "Mammography: Ducts as a sole indicator of breast carcinoma," Radiology **89**, 206–210 (1967).
[19] J. N. Wolfe, "The prominent duct pattern as an indicator of cancer risk," Oncology **23**, 149–158 (1969).
[20] J. N. Wolfe, "Breast patterns as an index of risk for developing breast cancer," Am. J. Roentgenol. **126**, 1130–1139 (1976).
[21] J. N. Wolfe, "Risk for breast cancer development determined by mammographic parenchymal pattern," Cancer (N.Y.) **37**, 2486–2492 (1976).
[22] I. E. Magnin, F. Cluzeau, C. L. Odet, and A. Bremond, "Mammographic texture analysis: An evaluation of risk for developing breast cancer," Opt. Eng. **25**, 780–784 (1986).
[23] J. W. Byng, N. F. Boyd, E. Fishell, R. A. Jong, and M. J. Yaffe, "Automated analysis of mammographic densities," Phys. Med. Biol. **41**, 909–923 (1996).
[24] J. W. Byng, N. F. Boyd, E. Fishell, R. A. Jong, and M. J. Yaffe, "The quantitative-analysis of mammographic densities," Phys. Med. Biol. **39**, 1629–1638 (1994).
[25] M. J. Yaffe, N. F. Boyd, J. W. Byng, R. A. Jong, R. Fishell, G. A. Lockwood, L. E. Little, and D. L. Tritchler, "Breast cancer risk and measured mammographic density," Eur. J. Cancer Prevention **7**, S47–S55 (1998).
[26] Z. Huo, M. L. Giger, D. E. Wolverton, and W. Zhong, "Computerized analysis of mammographic parenchymal patterns for breast cancer risk assessment: Feature selection," Med. Phys. **27**, 4–12 (2000).
[27] J. J. Heine and R. P. Velthuizen, "A statistical methodology for mammographic density detection," Med. Phys. **27**, 2644–2651 (2000).
[28] J. M. Boone, K. K. Lindfors, C. S. Veatty, and J. A. Seibert, "A breast density index for digital mammograms based on radiologists' ranking," J. Digit Imaging **11**, 101–115 (1998).
[29] J. N. Wolfe, A. F. Saftlas, and M. Salene, "Evaluation of mammographic densities: A case-control study," Am. J. Roentgenol., Radium Ther. Nucl. Med. **148**, 1087–1092 (1987).
[30] *American College of Radiology. Breast Imaging—Reporting and Data System (BI-RADS)*, 3rd ed. (American College of Radiology, Reston, VA, 1998).
[31] E. White, P. Velentgas, M. T. Mandelson, C. D. Lehman, J. G. Elmore, P. Porter, Y. Yasui, and S. H. Taplin, "Variation in mammographic breast density by time in menstrual cycle among women aged 40–49 years," J. Natl. Cancer Inst. **90**, 906–910 (1998).
[32] C. Zhou, H. P. Chan, N. Petrick, M. A. Helvie, M. M. Goodsitt, B. Sahiner, and L. M. Hadjiiski, "Computerized image analysis: Estimation of breast density on mammograms," Med. Phys. **28**, 1056–1069 (2001).
[33] S. J. Graham, M. J. Bronskill, J. W. Byng, M. J. Yaffe, and N. F. Boyd, "Quantitative correlation of breast tissue parameters using magnetic resonance and x-ray mammography," Br. J. Cancer **73**, 162–168 (1996).
[34] N. A. Lee, H. Rusinek, J. Weinreb, R. Chandra, H. Toth, C. Singer, and G. Newstead, "Fatty and fibroglandular tissue volumes in the breasts of women 20–83 years old: comparison of x-ray mammography and computer-assisted MR imaging," Am. J. Roentgenol. **168**, 501–506 (1997).
[35] B. Sahiner, N. Petrick, H. P. Chan, L. M. Hadjiiski, C. Paramagul, M. A. Helvie, and M. N. Gurcan, "Computer-aided characterization of mammographic masses: Accuracy of mass segmentation and its effects on characterization," IEEE Trans. Med. Imaging **20**, 1275–1284 (2001).
[36] B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images," IEEE Trans. Med. Imaging **15**, 598–610 (1996).
[37] O. Pawluczyk, B. J. Augustine, M. J. Yaffe, D. Rico, J. Yang, G. E. Mawdsley, and N. F. Boyd, "A volumetric method for estimation of breast density on digitized screen-film mammograms," Med. Phys. **30**, 352–364 (2003).

Design of Three-Class Classifiers in Computer-Aided Diagnosis: Monte Carlo Simulation Study

Heang-Ping Chan*, Berkman Sahiner, Lubomir M. Hadjiiski, Nicholas Petrick[a], Chuan Zhou
Department of Radiology, The University of Michigan, Ann Arbor, MI 48109
[a]Center for Devices and Radiological Health, U.S. Food and Drug Administration, Rockville, MD 20857

ABSTRACT

For the development of computer-aided diagnosis (CAD) systems, a classifier that can effectively differentiate more than two classes is often needed. For example, a detected object on an image may need to be classified as a malignant lesion, a benign lesion, or normal tissue. Currently, a three-class problem is usually treated as a two-stage, two-class problem, in which the detected object is first differentiated as a lesion or normal tissue, and, in the second stage, the lesion is further classified as malignant or benign. In this work, we explored methods for classification of an object into one of the three classes, and compared the three-class approach with the common two-class approach. We conducted Monte Carlo simulation studies to evaluate the dependence of the performance of 3-class classification schemes on design sample size and feature space configurations. A k-dimensional multivariate normal feature space with three classes having different means was assumed. Linear classifiers and artificial neural networks (ANNs) were examined. ROC analysis for the 3-class approach was explored under simplifying conditions. A performance index representing the normalized volume under the ROC surface (NVUS) was defined. Linear classifiers for classification of three classes and two classes were compared. We found that a 3-class approach with a linear classifier can achieve a higher NVUS than that of a 2-class approach. We further compared the performance of an ANN having three or one output nodes with a linear classifier. At large sample sizes, a 3-output-node ANN was basically the same as that of a one-output-node ANN. When the three class distributions had equal covariance matrices and the distances between pairs of class means were equal, the linear classifiers could reach a higher performance for the test samples than the ANN when the design sample size was small; the linear classifier and the ANNs approached the same performance in the limit of large design sample size. However, under complex feature space configurations such as the class means located along a line, the class in the middle was poorly differentiated from the other two classes by the linear classifiers for any dimensionality; the ANN outperformed the linear classifier at all design sample size studied. This simulation study may provide some useful information to guide the design of 3-class classifiers for various CAD applications.

KEY WORDS: Computer-aided diagnosis, classifier design, 3-class classification, linear classifier, artificial neural networks, Monte Carlo simulation, likelihood ratio, ROC analysis

1. INTRODUCTION

For the development of computer-aided diagnosis (CAD) systems, a classifier that can effectively differentiate more than two classes is often needed. For example, in an automated lesion detection and characterization system, it will be important to differentiate malignant lesions from benign lesions and normal tissue. A common approach is to treat this as a two-stage classification problem having two classes at each stage; masses are distinguished from normal tissue in the first stage, and then are classified as malignant and benign in the second stage. Alternatively, if the main interest is to detect only malignant lesions, a two-class classifier is trained to differentiate the malignant class from the combined class of the other two. The two classes that are treated as one may have very different characteristics and the classification may not be optimal if the classifier is forced to recognize their features as the same. For certain types of classification tasks, a properly designed 3-class classifier can be more effective in distinguishing one class from the other two classes. The design of 3-class classifiers has not been investigated systematically in the CAD area. In this work, we performed a simulation study to explore some properties of the 3-class and 2-class classification schemes.

* chanhp@umich.edu

2. MATERIALS AND METHODS

For an m-class classification problem in which a feature vector, **x**, is to be classified into one of m classes, a common approach is to apply the Bayes' rule to minimize the misclassification rate[1]. To accomplish this, the posterior probability of **x** belonging to class *i* is estimated as

$$p_i(c_i|\mathbf{x}) = g\, P(c_i)\, p(\mathbf{x}|c_i), \qquad \text{for } i = 1, \ldots, m \tag{1}$$

where $P(c_i)$ is the prior probability of class c_i, $p(\mathbf{x}|c_i)$ is the probability density of **x** in class c_i, and g is a constant. The feature vector is then assigned to class *k*, where *k* denotes the class that **x** has the maximum posterior probability,

$$k = \arg\max_{i=1,\ldots m}\{p_i(c_i \mid \mathbf{x})\} \tag{2}$$

However, it is difficult to estimate the posterior probability when the sample size is small. Furthermore, the misclassification rate does not take into account the fact that different types of misclassifications or correct classifications have different costs or utilities. A more general formulation of the m-class problem assigns a utility for each correct and incorrect decision, and optimizes the expected utility. The optimal decision rule depends on the utilities, as well as the prior probabilities of the classes. Let P_{ij} denote the probability of deciding class c_i when the true class is c_j, and U_{ij} denote the utility of deciding class c_i when the true class is c_j. The optimal decision rule is the one that maximizes the expected utility, which can be written as

$$E\{utility\} = \sum_{I=1}^{m}\sum_{j=1}^{m} U_{Ij} P_{Ij} P(c_j) \tag{3}$$

The limitation with the classifiers that maximize the correct classification rate or maximize the expected utility with fixed U_{Ij}'s is that they do not cover the entire range of sensitivity and specificity for the classification task. A receiver operating characteristic (ROC) analysis will provide the entire range of operating points. However, a 3-class classification problem will require a six-dimensional (6-D) ROC analysis as follows. For a 3-class problem with classes a (malignant), b (benign), and n (normal), there are nine possible "decision-truth" Ij pairs and hence nine probabilities: P_{Aa}, P_{Ba}, P_{Na}, P_{Ab}, P_{Bb}, P_{Nb}, P_{An}, P_{Bn}, P_{Nn}. Since the sum of every three of these probabilities is unity, e.g., $P_{Aa} + P_{Ba} + P_{Na} = 1$, only six of the nine probabilities are independent. Therefore, the ROC analysis will include these six possible variables.

For the 3-class problem, it has been shown that three decision lines that depend on two likelihood ratios (LRs) will provide the optimal decision boundaries on the LR plane as shown in Fig. 1[2] (C. E. Metz, private communication). The likelihood ratio between classes *i* and *j* is defined as the ratio of the probability density of **x** under each class,

$$LR_{ij}(\mathbf{x}) = p(\mathbf{x}|c_i)/\, p(\mathbf{x}|c_j) \tag{4}$$

In Fig. 1, the two LRs are chosen to be LR_{na} and LR_{ba}. The slopes and intercepts of the decision lines in the likelihood ratio plane depend on the prior probabilities of the classes, as well as the utilities of the different types of decisions, U_{Aa}, U_{Ba}, U_{Na}, U_{Ab}, U_{Bb}, U_{Nb}, U_{An}, U_{Bn}, U_{Nn}. The three decision lines always intersect at a common point. Varying the utilities and the priors over their allowed ranges will move the decision lines over the LR plane. For each configuration of the decision lines, the six probabilities can be estimated, producing a point in the 6-D ROC space. The complete treatment of a 6-D ROC analysis is therefore very complicated and has not yet been dealt with. In this study, we attempted to explore some properties of a 3-class problem under simplifying conditions.

We assume that the utilities can take on values in [0,1]. For correct decisions, the utilities will have the maximum value of 1, i.e., $U_{Aa} = U_{Bb} = U_{Nn} = 1$. If a malignant case is misdiagnosed as normal or benign, the utilities will be at a minimum of 0, $U_{Ba} = U_{Na} = 0$. If a normal case is called benign or vice versa, it may not be very harmful or costly so that the utilities $U_{Nb} = U_{Bn} = 1$. If a normal case or a benign case is called malignant, it will involve additional diagnostic tests or treatment and also cause patient anxiety or morbidity, the utilities U_{Ab} and U_{An} will be somewhere between 0 and 1. Under our assumptions that U_{Ab} and U_{An} are variable in (0,1) and the rest of the utilities are fixed as described above, it can be shown that two of the decision lines are reduced to one (Fig. 2), the third decision line becomes indeterminate, and the expected utility of the classification task in Eq. (3) depends only on three of the probabilities, P_{Aa}, P_{Ab}, and P_{An}. The 6-D ROC analysis will therefore be reduced to a 3-D ROC analysis under these

conditions. An example of the 3-D ROC surface is shown in Fig. 3. Note that P_{Aa} is the true-positive fraction (TPF) or the sensitivity, P_{Ab} is the false-positive fraction from the benign class (FPF_b), and P_{An} is the false-positive fraction from the normal class (FPF_n). This 3-D ROC surface is therefore similar to the commonly used 2-D ROC curve except that the FPF is split into the benign and normal classes. In analogy with the 2-D ROC analysis, we can define a performance index as the normalized volume under the 3-D ROC surface (NVUS) given by

$$\text{Normalized volume under 3D ROC surface (NVUS)} = \frac{\text{Volume under 3D ROC surface}}{\text{Projected area on the FP plane}} \tag{5}$$

Note that the NVUS can be interpreted as the average sensitivity over the range of FPF of interest, similar to the area under the 2-D ROC curve.

Ideally, if the feature vectors are transformed onto the LR plane, one can vary the decision line and determine the samples that fall into the region that is decided to be class A. The probabilities P_{Aa}, P_{Ab}, and P_{An} can then be estimated and the 3-D ROC surface generated. However, when the sample size is small, it is difficult to estimate the probability densities and derive the LRs for each **x**.

It is well-known that for the two-class classification problem in a k-dimensional feature space, the linear discriminant analysis projects the k-D feature space onto a 1-D decision axis. The decision boundary is then a threshold chosen along the decision axis. If the two class distributions are multivariate normal with equal covariance matrices, the linear discriminant classifier corresponds to the LR classifier and is optimal. This approach may be generalized to an m-class problem in a k-D feature space. In this case, the k-D feature space is projected to an (m-1)-D decision space, the decision boundaries are formed by (m-1) boundaries in the decision space[3]. For a 3-class problem (m=3), the k-D feature space is projected to a 2-D decision plane and the decision boundaries can be formed by two lines on the plane. In general, this projection is not optimal because it is not equivalent to a projection onto the LR plane. If the three class distributions are multivariate normal with equal covariance matrices, the linear transformation to a 2-D decision plane can be shown to be equivalent to a transformation to the log-likelihood ratio, *ln*(LR), plane and optimal decision boundaries can be formed on this plane.

In this preliminary study, we studied the 3-class classification problem by linearly projecting the k-D feature space to the 2-D decision plane and used two linear decision boundaries for differentiating the malignant class from the benign and the malignant classes. The classification performance was evaluated in the 3-D ROC space as shown in Fig. 3.

The 3-class classification was compared to the approach of treating the benign and normal classes as one (b+n) class such that the differentiation of the malignant class (class a) from the (b+n) class was considered to be a 2-class classification problem. The k-D feature space was thus projected to the 1-D decision line by linear discriminant analysis. This is equivalent to forming a hyperplane in the k-D feature space to separate class a from class (b+n).

We further assumed a simple k-D feature space in which the class distributions were multivariate normal, the covariance matrices for classes a, b, n were described by I, αI, αI, respectively, where I was the identity matrix and α was a constant. The mean vectors for the three classes were located at the vertices of an equilateral triangle. These characteristics are invariant upon projection to the 2-D decision plane in the 3-class classification approach described above although the scales may be changed. The 2-D decision plane in the 3-class classification approach shown in Fig. 4(a) and the example of the feature space in 2-D shown in Fig. 4(b) therefore have similar appearances. The symmetry of the class distributions about the vertical axis simplifies our analysis that follows, but the approaches should be applicable to non-symmetrical feature spaces.

For the 3-class classification approach, the slopes and intercepts of the linear decision boundaries were varied over the entire plane. For each set of boundaries, we could calculate the three probabilities, P_{Aa}, P_{Ab}, and P_{An} and generate a point in the 3-D ROC space. The surface formed by the highest sensitivity (P_{Aa}) at each FP location corresponded to the best decision boundaries. The NVUS was then derived from the highest sensitivity surface relative to its projected area on the FP plane.

For the 2-class classification approach with linear discriminant analysis, the best projection of the decision axis would be parallel to the symmetry (vertical) axis because of the symmetry of the class distributions. The decision boundary along this axis thus corresponded to a hyperplane perpendicular to the symmetry line. The decision boundary is illustrated as a horizontal line in the 2-D feature space (Fig. 4(b)). By moving the decision boundary along the decision axis and scoring the TPF and FPF, we could generate the 2-D ROC curve and derive the area under the ROC curve, A_z.

We compared the 3-class and 2-class approaches in two different ways. First, we compared the area under ROC curve under similar situations. For the 3-class approach and in our feature space with symmetry, the slice of the 3-D ROC surface along the diagonal of $P_{Ab} = P_{An}$ was equivalent to the situation of treating class b and class n equally, i.e., $U_{Ab} = U_{An}$. We calculated the area under the ROC curve obtained from this slice and compared it with the A_z obtained in the 2-class approach. In the second comparison, we modified the 2-class classification approach in the original k-D feature space. If we allowed the hyperplane to orient at an angle to the symmetry axis (the best projected decision axis in the linear discriminant analysis), it was similar to taking into consideration that there were different utilities of making FP decisions from class b or class n. For example, if the slope of the decision boundary was positive as shown in Fig. 5(a), we were less concerned with deciding a class-b sample as class a than deciding a class-n sample as class a so that it implied $U_{Ab} > U_{An}$. On the other hand, if the slope of the decision boundary was negative as shown in Fig. 5(b), we were less concerned with deciding a class-n sample as class a than deciding a class-b sample as class a so that it implied $U_{Ab} < U_{An}$. Therefore, by varying the slope and intercept of the single decision boundary in the 2-class approach, we could also generate a 3-D ROC surface and calculate its NVUS. We then compared the NVUS obtained from the 3-class and 2-class approaches.

Neural Network Classifiers

Another common approach that is often applied to the m-class classification problem is to use an artificial neural network (ANN) classifier with (m-1) output nodes. During training, the desired output of a sample from the i^{th} class is assigned 1 at the i^{th} node and assigned 0 at all other nodes. Under ideal conditions (sufficiently large training sample size and proper training), it has been shown that the ANN approaches a Bayes' classifier and the output for a given sample at the i^{th} node approaches the posterior probability of the sample in the i^{th} class[4]. Therefore, a properly trained ANN can be used for transforming the feature space to the LR plane and the 6-D ROC analysis applied. However, since the available design sample size is often limited in practice, the training of an ANN is usually far from being ideal. One of the common methods of analyzing the ANN output is to apply a 2-D ROC analysis to the scores of an individual output node, e.g., the i^{th} node, to distinguish the i^{th} class from the other classes. In this study, we evaluated the application of ANNs having one output node and three output nodes to the three-class problem. For training of the ANN with one output node, the desired output of the class-a samples was assigned to be 1 and those of the class-b and class-n samples were assigned to be 0. This is equivalent to treating the classification task as a 2-class problem without distinction between class b and class n. For training of the ANN with three output nodes, the desired output of a sample from the i^{th} class ($i = 1, 2, 3$) was assigned to be 1 at the i^{th} node and 0 at all other nodes. Under ideal conditions, one of the output nodes is actually redundant because the output of the third node is complementary to the other two. For both the 1-output-node ANN and the 3-output-node ANN, we applied 2-D ROC analysis to the output node that distinguished class a from the other two classes and compared the A_z values.

Simulation Study

We performed a simulation study to evaluate the different approaches discussed above. For this study, we assumed that the three class distributions were multivariate normal with diagonal covariance matrices. In a given experiment, 1000 samples were randomly drawn from the population for each of the classes. A subset of N_{train} trainers was randomly drawn from the 1000 available samples of each class and the rest, ($1000-N_{train}$), of the samples were held out as testers. N_{train} was varied from 20 to 600 per class for the linear classification study, and varied from 20 to 500 for the ANN study. For each condition, the experiment was repeated 50 times such that a new set of 1000 samples per class were drawn from the population. The dependence of the performance index, either A_z or NVUS, for each of the classification approaches on training sample size was evaluated. The ANNs were assumed to have one hidden layer with the number of hidden nodes equal to the number of input nodes. Backpropagation with a delta-bar-delta rule was used for training of the ANNs.

3. RESULTS

For comparison of the 3-class and 2-class approaches using linear classification, we assumed a 12-D multivariate normal feature space with covariance matrices *I*, 8*I*, 8*I* for class a, b, n, respectively. The comparison of A_z as a function of $1/N_{train}$ is plotted in Fig. 6. It can be seen that, for a given approach, when the design sample size is limited, the training (resubstitution) A_z is optimistically biased and the test (holdout) A_z is pessimistically biased, in comparison with the A_z at $N_{train} \to \infty$. The biases decrease as N_{train} increases. In the limit of $N_{train} \to \infty$, the training and test A_z approach essentially the same value. The A_z obtained from the 3-class approach is consistently higher than that from the 2-class approach for a given N_{train}.

Fig. 7 shows the comparison of the NVUS for the 3-class and 2-class approaches using linear classification in the same feature space. The characteristics of the curves are very similar to those observed in Fig. 6. The training NVUS is optimistically biased whereas the test NVUS is pessimistically biased compared to the limit achieved with large design sample size. The NVUS from the 3-class approach is again consistently higher than that from the 2-class approach for a given N_{train}.

For the comparison of the 3-output-node and 1-output-node ANNs, we first assumed a k-D (k=3, 6, 9, 12) multivariate normal feature space with equal covariance matrices *I*, *I*, *I* for class a, b, n, respectively. The dependence of A_z on $1/N_{train}$ is shown in Fig. 8(a) for the 3-output-node ANN and in Fig. 8(b) for the 1-output-node ANN. The characteristics of the A_z-versus-$1/N_{train}$ curves are very similar to those obtained in our previous study of 2-class classification problems[5]. The training A_z is optimistically biased and the test A_z is pessimistically biased compared with the A_z values at $N_{train} \to \infty$. The biases increase with the dimensionality of the feature space for a given N_{train} and decrease with increasing design sample size. It can be seen that the A_z values in the limit of $N_{train} \to \infty$ are very similar for the 3-output-node and the 1-output-node ANNs. For a given N_{train}, the biases of the 3-output-node ANN are larger than those of the 1-output-node ANN for the high dimensional feature spaces, probably because of the larger number of weights that need to be trained in the 3-output-node ANN with the finite design samples. For comparison, we also trained a linear discriminant classifier to differentiate class a from class (b+n) and plotted the A_z–versus-$1/N_{train}$ curves in Fig. 8(c). The A_z values in the limit of $N_{train} \to \infty$ from the linear classifiers are again very similar to those from the ANNs. These results indicate that the 3-output-node or the 1-output-node ANNs is basically performing 2-class classification at each of its output nodes. It is interesting to note that, when N_{train} is small, the biases in A_z from the linear classifier are much smaller than those from the ANNs. Therefore, in this feature space, when the design sample size is small, a linear classifier may be preferred over the ANNs because the performance of the trained linear classifiers is superior to that of the ANNs for unknown test samples.

The relative performance of the ANNs and linear classifiers depends strongly on the configuration of the class distributions, however. This can be demonstrated by comparing their performances in another multivariate normal feature space with unequal covariance matrices: class a had an identity matrix *I*, class b had a diagonal matrix with its diagonal elements varying from 1 to 2 in equal increment, class n had a diagonal matrix with its diagonal elements varying from 1 to 3 in equal increment. The three class means were lined up along a straight line in the k-D feature space. Fig. 9 shows an example of the class distributions in a 2-D feature space. The performances of the three classifiers in distinguishing class a, which is in the middle, from class b and class n are compared in Figs. 10(a) to 10(c). Under these conditions, the 3-output-node classifiers had slightly higher test A_z when N_{train} was small, but the A_z in the limit of $N_{train} \to \infty$ seemed to approach a level slightly lower than those of the 1-output-node ANN for the higher dimensional (9-D and 12-D) feature spaces. As expected, the linear classifiers were not able to distinguish the class a in the middle of class b and class n. Their performance was close to random guess for all sample sizes. These indicated that ANNs can be superior to a linear classifier for classification tasks with complex class distributions.

4. CONCLUSIONS

In this study, we explored some properties of 3-class and 2-class approaches to a 3-class classification task under simplifying conditions. By using Monte Carlo simulation study, we have examined the dependence of the performances of different classification schemes on design sample sizes for some feature space configurations. We

found that a 3-class approach can achieve higher classification accuracy than a 2-class approach under some conditions. Applying a 2-D ROC analysis to the output of a 3-output-node ANN achieved similar classification accuracy as that of a 1-output-node ANN. The ANNs may not be the method of choice for some classification tasks when the available design sample size is small. A complete treatment of 3-class classification using a 6-D ROC analysis is very complex and was not attempted in this preliminary study. Further investigation is underway to investigate if 3-class approaches can improve the accuracy for some classification tasks in CAD.

ACKNOWLEDGMENTS

This work is supported by USPHS grant CA95317 and U. S. Army Medical Research and Materiel Command Grants DAMD 17-02-1-0214. The content of this publication does not necessarily reflect the position of the funding agency, and no official endorsement of any equipment and product of any companies mentioned in this publication should be inferred. The authors are grateful to C. E. Metz, Ph.D., for his helpful discussion and notes on optimal decision variable.

REFERENCES

1. M. Nadler and E. P. Smith, *Pattern Recognition Engineering*, (John Wiley and Sons, New York, 1993).
2. H. L. Van Trees, *Detection, estimation, and modulation theory*, (John Wiley and Sons, New York, 1968).
3. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, (Wiley, New York, 1973).
4. C. M. Bishop, *Neural Networks for Pattern Recognition*, (Clarendon Press, Oxford, 1995).
5. H. P. Chan, B. Sahiner, R. F. Wagner and N. Petrick, "Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers," Medical Physics 26, 2654-2668 (1999).

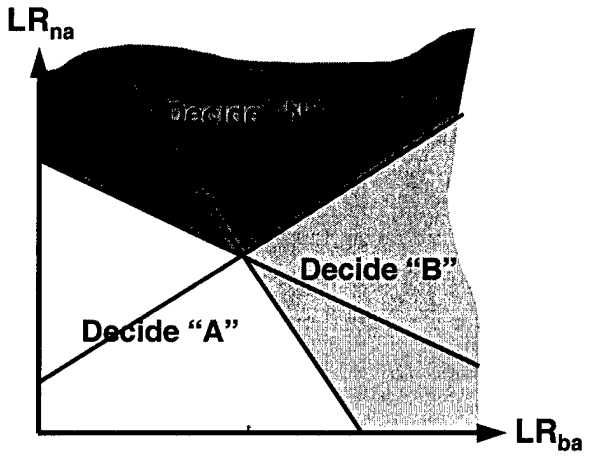


Fig. 1. Likelihood Ratio (LR) plane for a 3-class classification task.

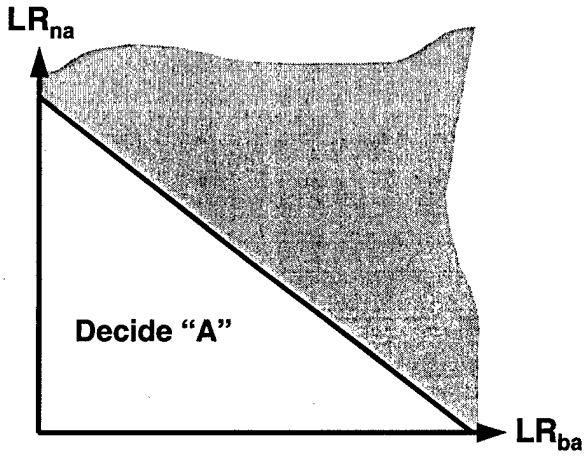


Fig. 2. Likelihood Ratio plane for a 3-class classification task under the assumptions in this study.

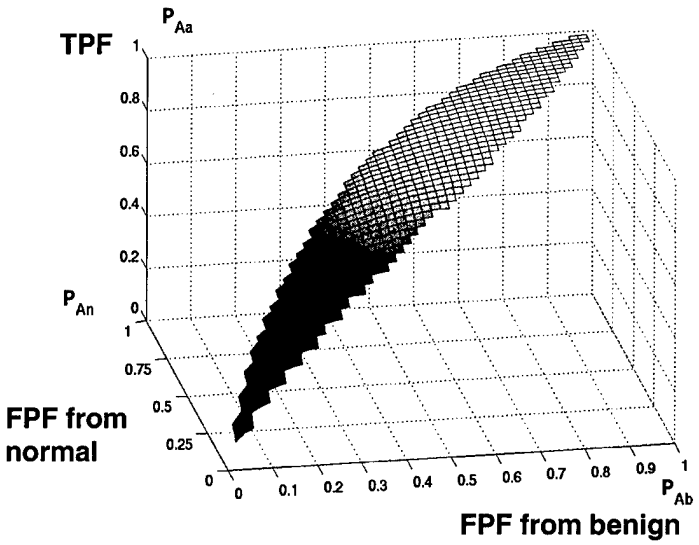


Fig. 3. 3-D ROC surface for the analysis of the 3-class classification problem under the assumptions in this study.

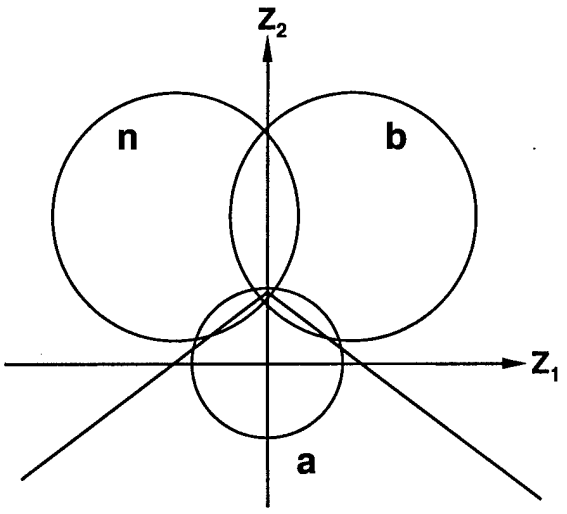


Fig. 4(a). Three-class approach for a 3-class classification task: 2-D decision plane with two linear decision boundaries.

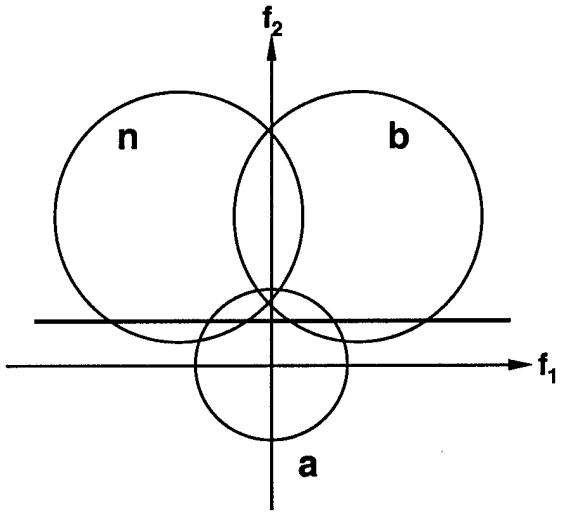


Fig. 4(b). Two-class approach for a 3-class classification task: k-D feature space (shown in 2-D as an example) with one linear decision boundary.

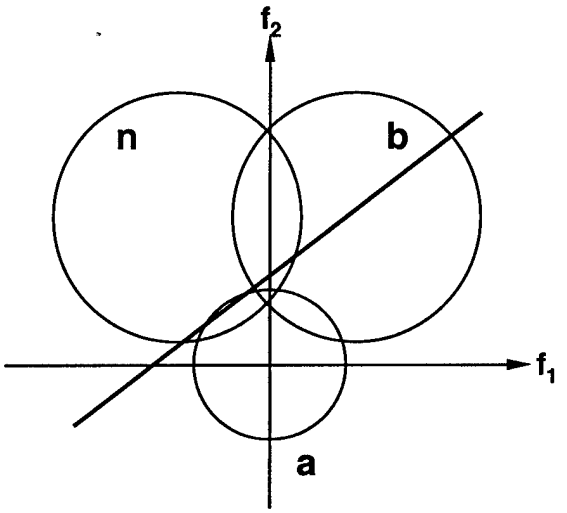


Fig. 5(a). Two-class approach for a 3-class classification task: a linear decision boundary that assumes $U_{Ab} > U_{An}$.

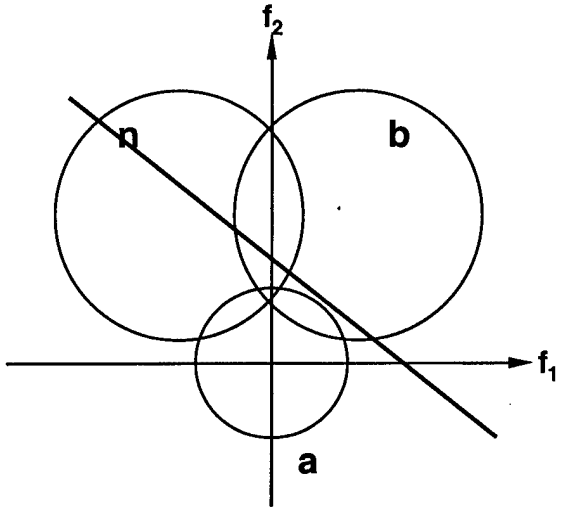


Fig. 5(b). Two-class approach for a 3-class classification task: a linear decision boundary that assumes $U_{Ab} < U_{An}$.

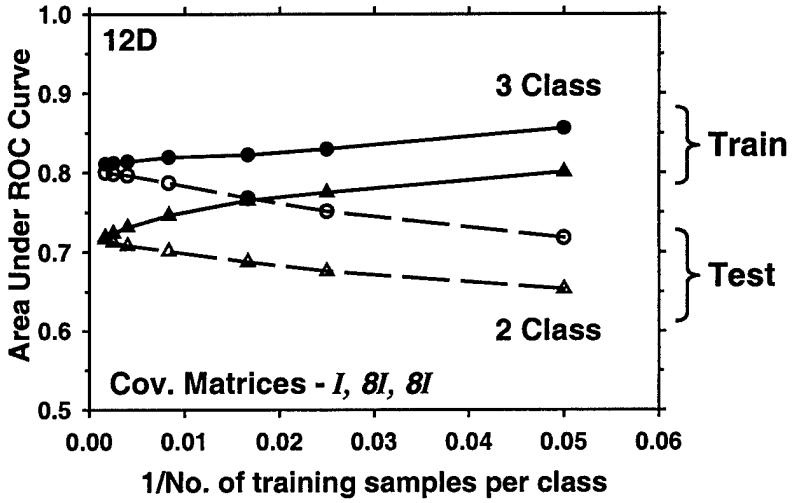


Fig. 6. Comparison of the performance of the 3-class and 2-class approaches for a 3-class problem. The areas under the 2-D ROC curves corresponding to $U_{Ab} = U_{An}$ are compared as a function of design sample size. Circles: 3-class approach. Triangles: 2-class approach. Solid curves: training results. Dashed curves: test results.

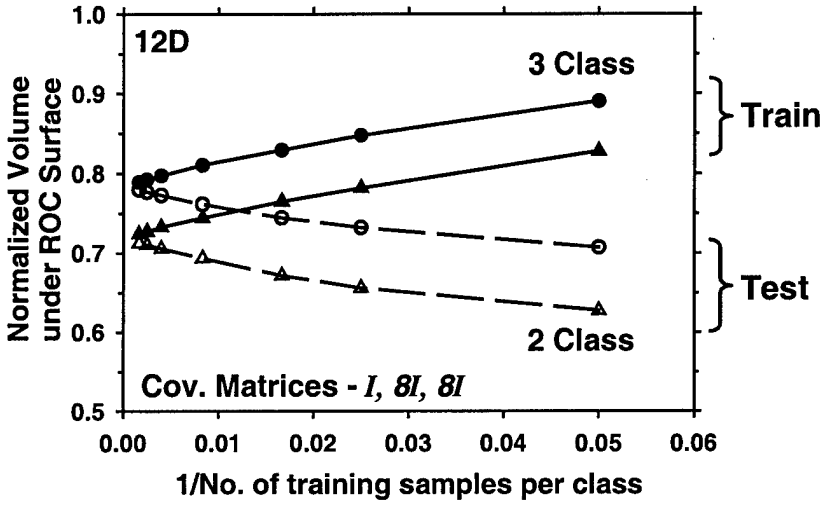
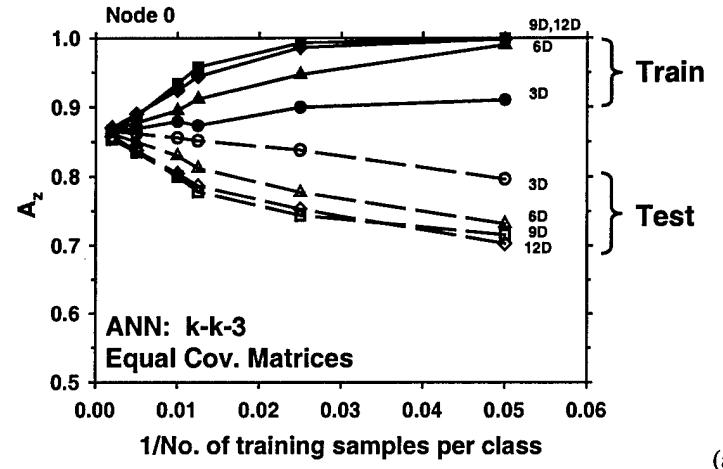
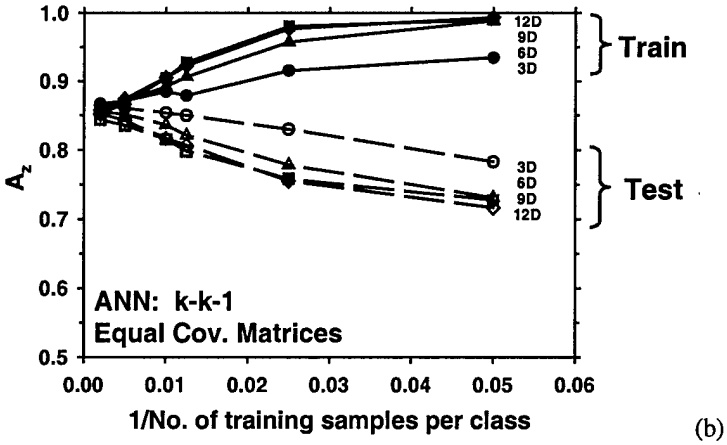


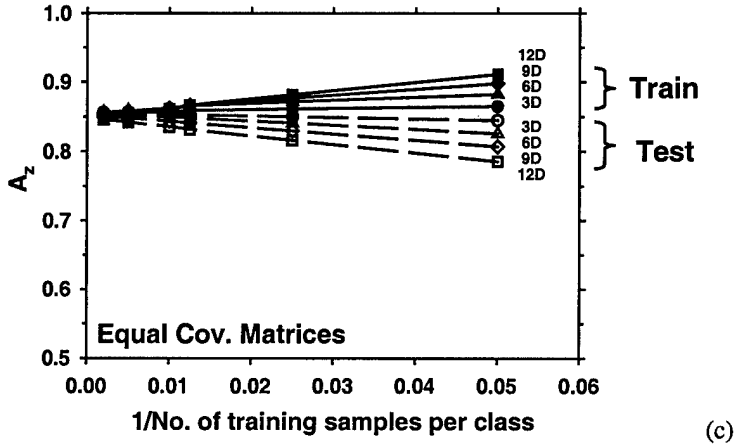
Fig. 7. Comparison of the performance of the 3-class and 2-class approaches for a 3-class problem. The normalized volumes under the 3-D ROC surface (NVUS) are compared as a function of design sample size. Circles: 3-class approach. Triangles: 2-class approach. Solid curves: training results. Dashed curves: test results.



(a)



(b)



(c)

Fig. 8. Classification performance in terms of A_z for differentiating class a from class b and class n. The class distributions in 3-D, 6-D, 9-D, 12-D feature spaces are multivariate normal with equal covariance matrices and class means located at the vertices of an equilateral triangle. (a) ANN: k input nodes, k hidden nodes, 3 output nodes, (b) ANN: k input nodes, k hidden nodes, 1 output node, and (c) linear classifier. Solid curves: training results. Dashed curves: test results.

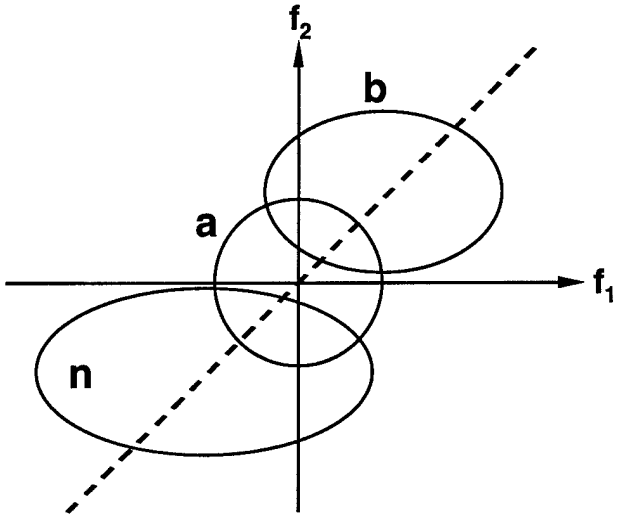
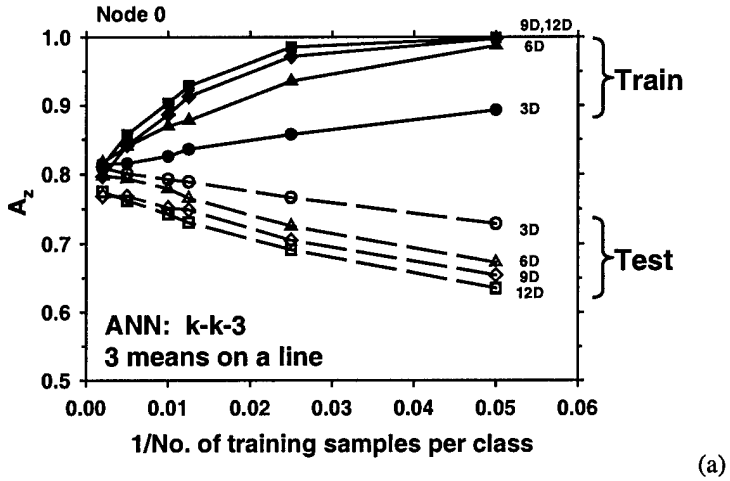
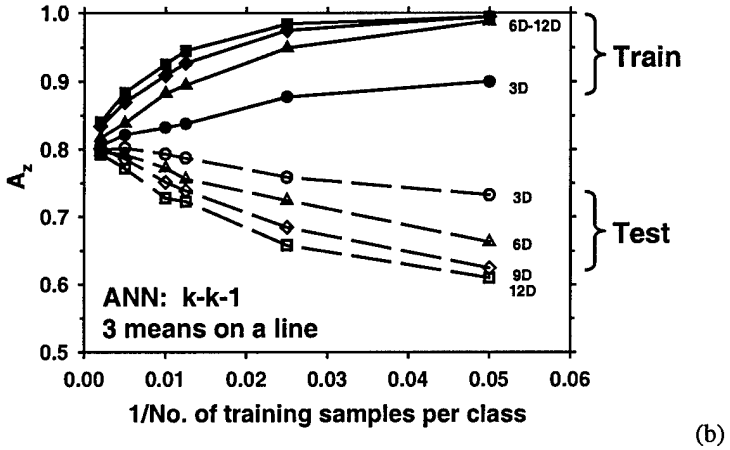


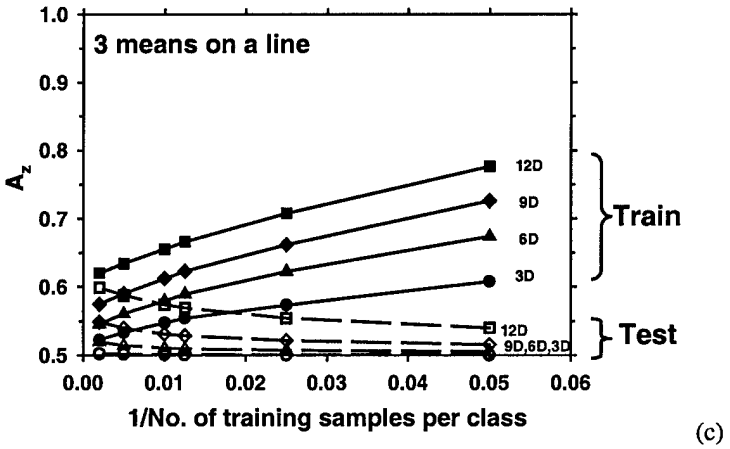
Fig. 9. A 3-class feature space with multivariate normal class distributions. The covariance matrices are diagonal and the three class means are located along a line. The example is illustrated in 2-D.



(a)



(b)



(c)

Fig. 10. Classification performance in terms of A_z for differentiating class a from class b and class n. The class distributions in 3-D, 6-D, 9-D, 12-D feature spaces are multivariate normal with unequal covariance matrices and class means along a line. (a) ANN: k input nodes, k hidden nodes, 3 output nodes, (b) ANN: k input nodes, k hidden nodes, 1 output node, and (c) linear classifier. Solid curves: training results. Dashed curves: test results.

ROC study: Effects of computer-aided diagnosis on radiologists' characterization of malignant and benign breast masses in temporal pairs of mammograms

Lubomir Hadjiiski*, Heang-Ping Chan, Berkman Sahiner, Mark A. Helvie, Marilyn Roubidoux, Caroline Blane, Chintana Paramagul, Nicholas Petrick[a], Janet Bailey, Katherine Klein, Michelle Foster, Stephanie Patterson, Dorit Adler, Alexis Nees, Joseph Shen

Department of Radiology, University of Michigan, Ann Arbor, MI; [a]Center for Devices and Radiological Health, U.S. Food and Drug Administration, Rockville, MD 20857

ABSTRACT

We conducted an observer performance study using receiver operating characteristic (ROC) methodology to evaluate the effects of computer-aided diagnosis (CAD) on radiologists' performance for characterization of masses on serial mammograms. The automated CAD system, previously developed in our laboratory, can classify masses as malignant or benign based on interval change information on serial mammograms. In this study, 126 temporal image pairs (73 malignant and 53 benign) from 52 patients containing masses on serial mammograms were used. The corresponding masses on each temporal pair were identified by an experienced radiologist and automatically segmented by the CAD program. Morphological, texture, and spiculation features of the mass on the current and the prior mammograms were extracted. The individual features and the difference between the corresponding current and prior features formed a multidimensional feature space. A subset of the most effective features that contained the current, prior, and interval change information was selected by a stepwise procedure and used as input predictor variables to a linear discriminant classifier in a leave-one-case-out training and testing resampling scheme. The linear discriminant classifier estimated the relative likelihood of malignancy of each mass. The classifier achieved a test A_z value of 0.87. For the ROC study, 4 MQSA radiologists and 1 breast imaging fellow assessed the masses on the temporal pairs and provided estimates of the likelihood of malignancy without and with CAD. The average A_z value for the likelihood of malignancy estimated by the radiologists was 0.79 without CAD and improved to 0.87 with CAD. The improvement was statistically significant (p=0.0003). This preliminary result indicated that CAD using interval change analysis can significantly improve radiologists' accuracy in classification of masses and thereby may increase the positive predictive value of mammography.

Keywords: Computer-Aided Diagnosis, Interval Changes, ROC Observer Study, Classification, Mammography, Breast Cancer.

1. INTRODUCTION

Mammography is currently the most sensitive method for detecting early breast cancer, and it is also the most practical screening exam [1,2] compared with other breast imaging techniques. However, the specificity of mammography is relatively low, only 15-30% of suspected breast lesions recommended for biopsy are actually malignant [3-5]. The unnecessary biopsies increase health care costs and cause patient anxiety and morbidity. If the specificity of differentiating malignant and benign mammographic lesions can be improved, the efficacy of mammography will be enhanced.

* L. H. (correspondence): e-mail:lhadjiski@umich.edu

Medical Imaging 2003: Image Processing, Milan Sonka, J. Michael Fitzpatrick, Editors, Proceedings of SPIE Vol. 5032 (2003) © 2003 SPIE · 1605-7422/03/$15.00

One of the important techniques that radiologists use in mammographic interpretation is to compare the current mammograms of a patient with those obtained in previous years, if available. The interval change information can help the detection of abnormalities, and identification of malignant breast lesions. It is shown that comparison with prior mammograms can improve both the sensitivity and specificity in breast cancer diagnosis [6,7].

In an early investigation, Chan et al. [8] demonstrated that computer-aided diagnosis (CAD) could improve significantly radiologists' detection of subtle mammographic microcalcification in an ROC study, This promising result stimulated continued development of CAD systems. To date, a number of CAD algorithms have been developed to detect suspicious masses and microcalcifications and to distinguish malignant and benign lesions on mammograms. Several ROC studies have shown that CAD systems could improve radiologists' accuracy in characterization of breast lesions. It has also been reported that CAD systems can increase the detection of breast cancers on screening mammograms in clinical practice[9,10].

Chan et al [11] performed an observer study to evaluate the effects of CAD, designed for characterization of malignant and benign masses on single view mammograms[12], on radiologists' diagnostic accuracy. They found that the radiologists' accuracy for classification of masses as malignant or benign in terms of the area under receiver operating characteristic (ROC) curve (A_z) was significantly improved ($p=0.022$ for one-view reading and 0.007 for two-view reading) with CAD compared to that without CAD. Huo et al [13] also conducted an observer study with 12 radiologists to classify masses on multiple views of mammograms. They also found that the radiologists' performance in terms of A_z was significantly improved ($p=0.001$) with computer aid. Jiang et al [14] performed an observer study to evaluate the effect of CAD on radiologists' classification of microcalcification clusters on mammograms. They found that with the computer aid the radiologists achieved a statistically significant improvement ($p<0.0001$).

The CAD systems for lesion classification so far employed information from a single exam.[12,14-19]. Based on the experiences of radiologists, it can be expected that even higher accuracy may be achieved if the computer can utilize the interval change information from multiple exams for classification. We recently[20] developed a classification scheme that combines prior and current information automatically extracted from masses on prior and current mammograms, respectively. We found that the classifier using the combined prior and current information performed significantly better ($p=0.015$) in terms of A_z than the classifier using current information alone. The current study investigated the effects of CAD on assisting radiologists in evaluating interval changes in serial mammograms. To our knowledge, this is the first ROC experiment to evaluate the effects of a computer classifier using interval change information on radiologists' diagnosis of breast cancers.

2. MATERIALS AND METHODS

2.1 Data set

We selected a set of 126 temporal pairs of mammograms containing biopsy-proven masses on the current mammograms from our database. The mammograms in the database were digitized consecutively from the patients who had undergone breast biopsy in our department. The selection criterion used in the current study was that the case had serial exams in which a corresponding mass could be identified. The mammograms thus contained masses covering a range of sizes and conspicuity that will be seen in clinical practice. The data set consisted of 220 mammograms from 52 patients. The mammograms were digitized with a LUMISCAN 85 laser scanner at a pixel resolution of $50\mu m \times 50\mu m$ and 4096 gray levels. The image matrix size was reduced by averaging every 2 x 2 adjacent pixels and down-sampled by a factor of 2 to obtain images with a pixel size of $100\mu m \times 100\mu m$ for analysis by the computer.

There were 53 biopsy proven masses (32 malignant and 21 benign) in the 52 cases. The 220 mammograms contained different mammographic views (CC, MLO, and lateral views) and multiple serial examinations of the masses including the examination when the biopsy decision was made. By matching masses of the same view from two different examinations, a total of 126 temporal pairs were formed, of which 73 were malignant and 53 benign. Since all cases in this data set had undergone biopsy, the benign masses in this set could not be distinguished easily from the malignant ones based on current mammographic criteria. For the malignant masses in this data set, the average mass size was 7.9 mm on the prior mammograms and 12.0 mm on the current mammograms. The corresponding sizes were 9.8 mm and 11.4 mm, respectively, for the benign masses.

To simulate a more realistic clinical situation 34 additional temporal pairs containing corresponding normal structures in the serial mammograms were also included. In this way the radiologist also has to distinguish mass-mimicking fibroglandular tissue from malignant masses. The temporal pairs had a time interval of 6 to 48 months. More than 67% of the pairs had a time interval of 12 months.

2.2 Design of classifier for classification of masses in serial mammograms

We have developed a novel classification technique that utilizes the current and prior information on serial mammograms to characterize the masses. The classification technique has been described in detail elsewhere[20]. The method is summarized in the flowchart shown in Figure 1. Initially a region of interest (ROI) containing the mass was defined by a radiologist on both the current and prior mammograms. Automatic segmentation of the mass within each ROI was performed based on an active contour model [21,22]. A set of texture, morphological, and spiculation features were extracted for each mass.

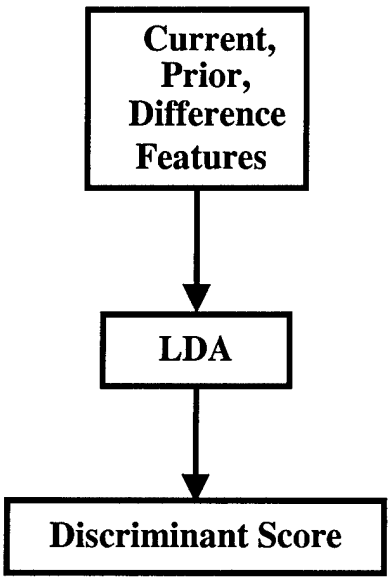


Figure 1. Block-diagram of the classification method.

The texture features were based on run-length statistics (RLS) matrices [23]. The RLS matrices were computed from the images obtained by the rubber band straightening transform (RBST) [12]. The RBST maps a band of pixels surrounding the mass onto a rectangular region. Five texture measures were extracted from the vertical and horizontal gradient images derived from the RBST image in two directions [12]. Therefore, for each ROI, a total of 20 RLS features were calculated. Morphological features were extracted from the automatically segmented mass shape and gray levels [22,24]. Spiculation features were extracted by using the statistics of the image gradient direction relative to the normal direction to the mass border in a ring of pixels surrounding the mass [21,22]. A total of 35 features (20 RLS, 12 morphological and 3 spiculation) were therefore extracted from each ROI. Additionally, difference features were obtained by subtracting a prior feature from the corresponding current feature, resulting in 35 difference features.

A "leave-one-case-out" resampling scheme was used for the training and testing of the classifier. In order to reduce the dimensionality of the feature space, a stepwise feature selection was employed to select the most effective

feature subset from each training cycle. An average of 7 features were selected for the classification task from the training subsets.

A relative malignancy rating by the computer classifier on a scale of 1 to 10 was provided to the radiologists for the reading with CAD. The relative malignancy rating was obtained by linearly scaling the classifier output within the interval between 1 and 10 and then rounding the result to the closest integer. A higher rating corresponded to a higher likelihood of being malignant. Gaussian functions were fitted to the distributions of the malignant and benign samples to obtain a fitted binormal distribution with the classifier's malignancy ratings scaled to the range of 1 to 10 (Figure 2). The fitted distribution was displayed on the graphical user interface as a reference when the radiologist evaluated the cases using CAD.

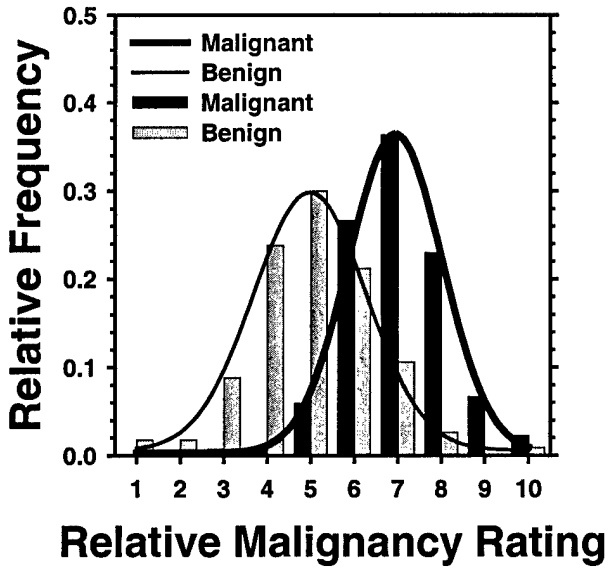


Figure 2. Binormal distribution fitted to the histogram.

2.3 Radiologist's classification of masses in serial mammograms

The observer study was designed to compare radiologists' performance on the classification of malignant and benign breast masses with and without CAD. The ROIs extracted from the current and the prior mammograms containing the corresponding mass was displayed side-by-side on a display monitor. The observers' performance was evaluated under two reading conditions. In the first reading condition, the radiologist read the temporal image pair of the mass without computer aid. In the second reading condition, the radiologist read the temporal pair with computer classifier's relative malignancy rating of the mass displayed on the screen. The observer was asked to provide an estimate of the likelihood of malignancy of the mass in a 100-point rating scale under each reading condition. Four MQSA radiologists and one breast imaging fellows participated as observers in this study.

A counter-balanced design was used in arranging the reading orders in different modes and the case orders in different reading sessions for the observers. This approach would minimize the potential effects such as learning,

fatigue, and memorization on the outcomes of the observer experiments. A graphic user interface was developed for the purpose of presenting the temporal pairs of mass ROIs to the radiologists and recording their ratings. Each observer underwent a training session before the actual reading sessions to familiarize them with the performance of the CAD system and the experimental procedure.

2.4 ROC analysis

The likelihood of malignancy ratings of the individual observers for the two reading conditions were analyzed by using ROC methodology. A binormal ROC curve was fitted to each observer's 100-point rating data for each reading condition by the LABROC program using maximum likelihood estimation.[25] The classification accuracy was quantified by using the total area under the fitted ROC curve, A_z,. The slope and the intercept parameters for the individual ROC curves were also estimated by the LABROC program. For each reading condition, the average performance of the radiologists was estimated as the area under an average ROC curve, which was derived from the average slope and intercept parameters of the 5 individual observer's ROC curves for that reading condition. The statistical significance of the difference in A_z between the two reading conditions was estimated by the Student's two-tailed paired t-test on the 5 pairs of individual observer's A_z values.

3. RESULTS

The A_z values for the 5 radiologists participating in the study for the two reading conditions with and without CAD are presented in Fig 3. The computer classifier's test A_z value was 0.87. The average ROC curves for the 5 observers when reading with and without CAD were plotted in Fig.4. The A_z value from the average ROC curve was 0.79 for reading without CAD and 0.87 for reading with CAD. The radiologist performance was improved, both individually and on average, when reading with the CAD system. The improvement in the average A_z between the reading without CAD and the reading with CAD was statistically significant (Student's two-tailed paired t-test, p=0.0003).

The computer classifier's A_z value of 0.87 was higher than the individual radiologists' A_z values obtained under the reading condition without CAD. The relatively low accuracy of the radiologists in classifying the masses reflected the fact that these were difficult cases that all had been recommended for biopsy. All five radiologists improved their accuracy in classification of the malignant and benign masses when the CAD system was available as a second opinion. Two radiologists achieved an A_z higher than that of the computer classifier under the reading condition with CAD. We did not observe specific differences between the breast imaging fellow compared to the MQSA-approved radiologists. The improvement in A_z ranged between 0.06 and 0.1.

4. CONCLUSION

We have performed an observer ROC study to evaluate the effects of computer-aided diagnosis on radiologists' characterization of masses on serial mammograms. In this observer study the radiologists improved their performance with statistically significance ($p = 0.0003$) when their reading without computer aid was compared to that with computer aid. These results suggest that CAD may be helpful in improving the accuracy of malignant and benign mass characterization.

ACKNOWLEDGMENTS

This work was supported by USAMRMC grants DAMD17-98-1-8211, DAMD17-02-1-0489, and DAMD17-02-1-0214. The authors are grateful to Charles E. Metz, Ph.D., for the LABROC program.

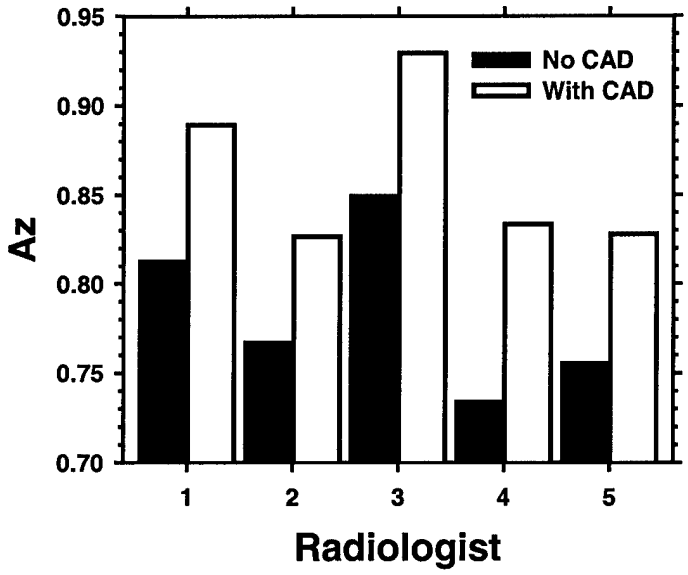


Figure 3. The area under ROC curve, A_z, for the characterization of the masses in 126 pairs of serial mammograms by 5 radiologists under two reading conditions: without CAD and with CAD. The average A_z for the two reading conditions: no CAD (A_z=0.79), with CAD (A_z=0.87).

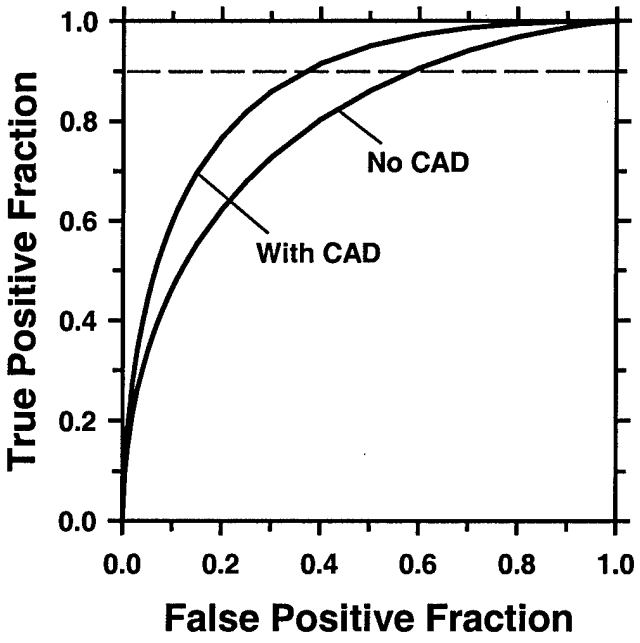


Figure 4. Area under ROC curve for the mode without CAD and the mode with CAD by the 5 radiologists. Average area for the two reading modes: No CAD (A_z=0.79), With CAD (A_z=0.87). The difference is statistically significant (Student paired t-test, p=0.0003).

REFERENCES

1. H. C. Zuckerman, "The role of mammography in the diagnosis of breast cancer," *In: Breast Cancer, Diagnosis and Treatment*, 152-172, Eds. I. M. Ariel and J. B. Cleary, McGraw-Hill, New York, 1987.

2. L. Tabar and P. B. Dean, "The Control of Breast Cancer through Mammography Screening," *Radiologic Clincs of North America* **25**, 961, 1987.

3. E. A. Sickles, "Mammographic features of 300 consecutive nonpalpable breast cancers," *American Journal of Roentgenology* **146**, 661-663, 1986.

4. D. B. Kopans, "The positive predictive value of mammography," *American Journal of Roentgenology* **158**, 521-526, 1991.

5. D. D. Adler and M. A. Helvie, "Mammographic biopsy recommendations," *Current Opinion in Radiology* **4**, 123-129, 1992.

6. L. W. Bassett, B. Shayestehfar, and I. Hirbawi, "Obtaining previous mammograms for comparison: usefullness and costs," *Amer. J. Roentgenology* **163**, 1083-1086, 1994.

7. E. A. Sickles, "Periodic mammographic follow-up of probably benign lesions: results in 3183 consecutive cases," *Radiology* **179**, 463-468, 1991.

8. H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis," *Investigative Radiology* **25**, 1102-1110, 1990.

9. M. A. Helvie, L. M. Hadjiiski, E. Makariou, H. P. Chan, N. Petrick, and S. B. Lo, "A Non-Commercial CAD System for Breast Cancer Detection on Screening Mammograms Achieves High Sensitivity : A Pilot Clinical Trial," *Radiology* **225(P)**, 459, 2002.

10. T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center," *Radiology* **220**, 781-786, 2001.

11. H.-P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman, and S. S. Gopal, "Improvement of radiologists' characterization of mammographic masses by computer-aided diagnosis: an ROC study," *Radiology* **212**, 817-827, 1999.

12. B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," *Medical Physics* **25**, 516-526, 1998.

13. Z. M. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Breast Cancer: Effectiveness of Computer-aided Diagnosis - Observer Study with Independent Database of Mammograms," *Radiology* **224**, 560-568, 2002.

14. Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," *Academic Radiology* **6**, 22-33, 1999.

15. H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Leung, D. D. Adler, and M. M. Goodsitt, "Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network," *Physics in Medicine and Biology* **42**, 549-567, 1997.

16. Z. M. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Academic Radiology* **5**, 155-168, 1998.

17. J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computer-aided image analysis," *IEEE Transactions on Medical Imaging* **12**, 664-669, 1993.

18. L. M. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, and M. A. Helvie, "Classification of malignant and benign masses based on hybrid ART2LDA approach," *IEEE Transactions on Medical Imaging* **18**, 1178-1187, 1999.

19. G. D. Tourassi, M. K. Markey, J. Y. Lo, and C. E. Floyd, "A neural network approach to breast cancer diagnosis as a constraint satisfaction problem," *Medical Physics* **28**, 804-811, 2001.

20. L. M. Hadjiiski, B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. N. Gurcan, "Analysis of Temporal Change of Mammographic Features: Computer-Aided Classification of Malignant and Benign Breast Masses," *Medical Physics* **28**, 2309-2317, 2001.

21. B. Sahiner, H. P. Chan, N. Petrick, L. M. Hadjiiski, M. A. Helvie, and S. Paquerault, "Active contour models for segmentation and characterization of mammographic masses," *The 5th International Workshop on Digital Mammography*, 357-362, Toronto, Canada, 2001.

22. B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, and L. M. Hadjiiski, "Improvement of mammographic mass characterization using spiculation measures and morphological features," *Medical Physics* **28**, 1455-1465, 2001.

23. M. M. Galloway, "Texture classification using gray level run lengths," *Computer Graphics and Image Processing* **4**, 172-179, 1975.

24. N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms," *Medical Physics* **26**, 1642-1654, 1999.

25. D. D. Dorfman, K. S. Berbaum, and C. E. Metz, "ROC rating analysis: Generalization to the population of readers and cases with the jackknife method," *Investigative Radiology* **27**, 723-731, 1992.